COMPUTATIONAL METHODS TO STUDY GENE REGULATION IN HUMANS USING DNA AND RNA SEQUENCING DATA

by

Ashis Saha

A dissertation submitted to Johns Hopkins University
in conformity with the requirements for the degree of
Doctor of Philosophy

Baltimore, Maryland
March 2021

Abstract

Genes work in a coordinated fashion to perform complex functions. Disruption of gene regulatory programs can result in disease, highlighting the importance of understanding them. We can leverage large-scale DNA and RNA sequencing data to decipher gene regulatory relationships in humans. In this thesis, we present three projects on regulation of gene expression by other genes and by genetic variants using two computational frameworks: co-expression networks and expression quantitative trait loci (eQTL).

First, we investigate the effect of alignment errors in RNA sequencing on detecting trans-eQTLs and co-expression of genes. We demonstrate that misalignment due to sequence similarity between genes may result in over 75% false positives in a standard trans-eQTL analysis. It produces a higher than background fraction of potential false positives in a conventional co-expression study too. These false-positive associations are likely to misleadingly replicate between studies. We present a metric, *cross-mappability*, to detect and avoid such false positives.

Next, we focus on joint regulation of transcription and splicing in humans. We present a framework called *transcriptome-wide networks (TWNs)* for combining total expression of genes and relative isoform levels into a single

sparse network, capturing the interplay between the regulation of splicing and transcription. We build TWNs for 16 human tissues and show that the hubs with multiple isoform neighbors in these networks are candidate alternative splicing regulators. Then, we study the tissue-specificity of network edges. Using these networks, we detect 20 genetic variants with distant regulatory impacts.

Finally, we present a novel network inference method, *SPICE*, to study the regulation of transcription. Using maximum spanning trees, SPICE prioritizes potential direct regulatory relationships between genes. We also formulate a comprehensive set of metrics using biological data to establish a standard to evaluate biological networks. According to most of these metrics, SPICE performs better than current popular network inference methods when applied to RNA-sequencing data from diverse human tissues.

Thesis Readers

Alexis Battle (Advisor)
Associate Professor
Department of Biomedical Engineering and
Department of Computer Science
Johns Hopkins University

Benjamin Langmead
Associate Professor
Department of Computer Science
Johns Hopkins University

Michael Schatz
Bloomberg Distinguished Associate Professor
Department of Computer Science and
Department of Biology
Johns Hopkins University

For my family – the biggest source of happiness.

Arjun Kumar Saha	*Babu* knows how to inspire me.
Lila Bati Podder	*Maa* sacrifices a lot to fulfill my wishes.
Asim Kumar Saha	My idol.
Snigdha Saha	Inseparable connection.
Jawshan Ara	The love of my life.

Acknowledgments

I am deeply grateful to my advisor Alexis Battle for her support and supervision. She provided me with a perfect balance between freedom and guidance: she encouraged me to be independent, always supported my plans, and guided me to reach my goals. I could not have asked for a better mentor. The more I hear about others' advisors, the more fortunate I feel for having such an amazing adviser. I continue to be amazed by her keen intellect, integrity, and kindness. She is my inspiration. Thank you, Alexis, for everything!

I am thankful to Ben Langmead who helped me grow since the beginning of my graduate life. In particular, he patiently supervised one of my qualifying projects. He was also on my GBO committee. He gave valuable comments and suggestions on other projects too. He is an excellent teacher and researcher; I learned a lot from him. Thank you, Ben, for believing in me. I am also thankful to Mike Schatz for his support and being on my GBO and thesis committee. I thank Jeff Leek, James Taylor, Steven Salzberg, and Kasper Hansen for agreeing to be on my GBO committee.

It has been a pleasure to work closely with outstanding collaborators including Yungil Kim, Barbara E. Engelhardt, Ariel Gewirtz, Brian Jo, Chuan Gao, Ian McDowell, Daniel Geschwind, Kasper Lage, Chris Hartl, Sandrine

Muller, Gokul Ramaswami, Michael Gandal, William Pembroke, François Aguet, Christopher Brown, Stephen Montgomery, Kristin Ardlie, Tuuli Lappalainen. I learned a tremendous amount from these collaborations. I am grateful to the members of the GTEx consortium; it was a privilege to collaborate with world-class researchers. The GTEx consortium taught me the importance of teamwork to achieve extra-ordinary goals.

I have been fortunate to be surrounded by a group of talented, smart, intelligent, and yet humble and friendly folks – my labmates. With them, I have enjoyed long academic discussions a lot. I have enjoyed non-academic interactions (e.g., coffee breaks, lunch/dinner, sharing thoughts about life, walks, etc.) even more. They helped and supported me to grow both as a researcher and as a human being. Thank you - Princy Parsana, Yungil Kim, Yuan He, Surya Chhetri, Rebecca Keener, Matthew Figdore, Ben Strober, Prashanthi Ravichandran, Benj Shapiro, Zeinab Mousavi, Diptavo Dutta, Marios Arvanitis, Josh Popp, Ashton Omdahl, Taibo Li, April Kim, Guanghao Qi, Farhan Damani, Karl Tayeb, Amy He, Bohan Ni, Boris M Brenerman, Jessica Bonnie, Victor Wang, Milind Agarwal, Seraj Grimes, Eric Kernfeld, and Afrooz Razi. I always appreciate your kindness and friendship.

I feel lucky to get a cooperative and collaborative research environment at Hopkins. In particular, bi-weekly joint genomics meetings helped me know about excellent research happening around me. I also received valuable comments about my research in these meetings. Thanks to the organizers. Special thanks to Jacob Pritt, Christopher Wilks, Abhinav Nellore, Leonardo Collado-Torres, and Johann Hawe for sharing their knowledge and thoughts.

I am thankful to my colleagues and friends Anand Malpani, Charlotte Darby, Daniel Baker, Taher Mun, and Sam Kovaka. Thanks to wonderful staff members at Hopkins who helped me during my whole graduate life: Sarah Anderson, Deborah DeFord, Zachary Burwell, Cathy Thornton, Kim Franklin, Laura Graham, Tonette McClamy, Shani McPherson, Steve Rifkin, Steve DeBlasio, and Christopher Venghaus.

I have been blessed with many near and dear friends without whom life abroad could have been difficult. They have become my extended family away from home. Jayanti and Gautom have been a constant support system from early days in the USA to the current pandemic period. I got sisters and brothers in Zajeba, Shaolin, Manashi, Pial, Nafi, Uzma, Mukib, Tanna, Tanvir, Ema, Shehab, Silvia, Shila, Salaheen, Adnan, Zinat, Saad, Mouri, Sayma, Chayan, Sati, Rayhan, Trisha, Amit, Jibon, Uday, Palash, and Jayanta. We spent time together, arranged cultural programs together, played together, traveled together, and celebrated together. Reuniting with good friends after a long time has been a highlight of my life abroad. Uttam, Laboni, Atif, Ove, Liza, Dollar, Meghna, Shahid, Putul, Amit, Ashis, Suman, Nandini, Sajib, and Shante are some of the highlights. My *Badhan* friends in the USA – Sohel, Snigdha, Imran, Adnan, Shubhra, Brinta, Faruque, Shanta, and Nodi – made me feel home. Hasan, Swakkhar, Iftee, Shyamal, Sabiranjan, and Sohag were as if living nearby. Occasional visits to and from Minakshi, Govinda, Rony, Shamim, Shamsul, and Lipi brought happiness to our life. The next generation – Riddho, Aneesha, Kolpo, Rhiju, Aarjo, Rujul, Rushan, Rafin, and Manha – especially made our life colorful. Thank you all (including whom I missed)

for being a part of our life; I could not have survived without you.

I am immensely grateful to my parents – Arjun Kumar Saha and Lila Bati Podder – for their unconditional love, support, and guidance. During childhood, my dad introduced me to the beauty of mathematics using real-life examples that formed the foundation of my scientific career. My mom is a living example of the importance of an educated working woman in a family. My siblings and I could come this far because of her foresight and her support. She sacrificed a lot for us to fulfill our wishes, to make us happy. I cannot thank my parents enough. I must thank my siblings Asim Kumar Saha and Snigdha Saha. Asim has been my idol ever since I could remember; he has been my go-to person for everything. I spent most of my childhood with Snigdha, fought with her a lot. We care a lot for each other – our connection is inseparable. I am grateful to my aunt Kanika Podder and uncle Sunil Chandra Chowdhury without whom quality education for our siblings could have been difficult. Thanks to my sister-in-law Aparna Saha, brother-in-law Anup Kumar, sister-in-law Nawshin, mother-in-law Lutfun Nahar, father-in-law Md Badrul Islam, and cousins Susmita, Tumpa, Manas, Chaity, Pulak, Loknath, and Hridoy. Special thanks to my sweet nephews and nieces - Ahona, Mounota, Purnota, Orish, and Progga.

Last but not least, I am grateful to my wife and partner in crime Jawshan Ara Shatil. We have crossed a world, both literally and metaphorically, to stay together. Our journey is just beginning. Thank you, Shatil, for your love, support, and inspiration – I could not have completed my Ph.D. without you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

We can view biology, the science of life and living organisms, as the study of interconnections within and across multiple layers: from atoms to cells to tissues to organisms (Craig et al., 2014). Each layer performs certain functions in coordination with other layers. It is necessary to store and transport information for such coordinated, often complex, tasks. Primarily, cells in a living organism use DNA (deoxyribonucleic acid) for long-term information storage and RNA (ribonucleic acid) for information transport. According to the central dogma of life, DNA is first *transcribed* into RNA which in turn is *translated* into protein. Segments of DNA that code for RNA or protein are called *genes*. The process to synthesize RNA or protein from a gene is called *gene expression*. Disruption of gene regulatory programs often results in critical diseases (Lee & Young, 2013), highlighting the importance of understanding them.

Laboratory experiments traditionally have been the dominating factor in understanding biology, successfully elucidating fundamental molecular principles for centuries. Each experiment generally produces some data which

is analyzed by humans to gather support for or against a hypothesis. With the advent of recent technologies, the dimension of certain types of biological data has become so big that scientists can no longer analyze them manually. Consequently, large scale computational and statistical methods are becoming increasingly popular day by day (Biotechnology, 2016). These methods not only increase the computational power, but also widen the hypothesis space. By appropriate modelling, we may even answer questions for which the experiments were not initially designed. Perhaps more importantly, we can attempt to test a hypothesis where a proper experiment cannot be performed due to ethical, economical or other reasons. For example, while we may not perform a gene knockout experiment in humans to test the function of a gene due to ethical reasons, we can apply computational methods to predict the function.

Once we know about gene functions and their regulatory mechanisms, we can utilize the knowledge to diagnose, prevent, and treat diseases. Recent biotechnological advancements are revolutionizing medical sciences through vaccine design, drug development, gene therapy, personalize medicine, and other areas. In fact, more than 250 genomic biomarker-based drugs have already been approved by the U.S. Food and Drug Administration (FDA) (U.S. Food and Drug Administration, 2020).

In this thesis, we focus on computational methods to study gene regulation in humans. Using DNA and RNA sequencing data, we focus on how the expression of a gene is regulated by either other genes or genetic variants. We hope the knowledge acquired here would ultimately contribute to improve

human health.

1.1 Background

Before diving into the thesis, we provide a brief overview of a few basic concepts in molecular biology that form the basis of the thesis. For simplicity, we focus only on typical scenarios in humans. But, please be aware that there might be exceptions.

1.1.1 Cells and chromosomes

The *cell* is the structural and functional unit of all living organisms. A typical human adult has about 37.2 trillion cells (Bianconi et al., 2013). There is a *nucleus* in each human cell. The nucleus contains all the chromosomes. Each *chromosome* – a threadlike structure of nucleic acids and protein – is actually a tightly-packaged DNA molecule that carries genetic information of the organism (see Figure 1.1). Humans have 23 pairs of chromosomes: 22 pairs of autosomes (chr1, chr2, chr3, . . . , chr22) and one pair of sex chromosomes (chrX and chrY).

1.1.2 DNA

DNA (deoxyribonucleic acid) molecules carry the hereditary blueprint of humans and most other organisms. A DNA molecule is composed of two polynucleotide strands each coiled around the same axis to form a double-helix structure (Watson & Crick, 1953). Each strand is a long chain of *nucleotides*,

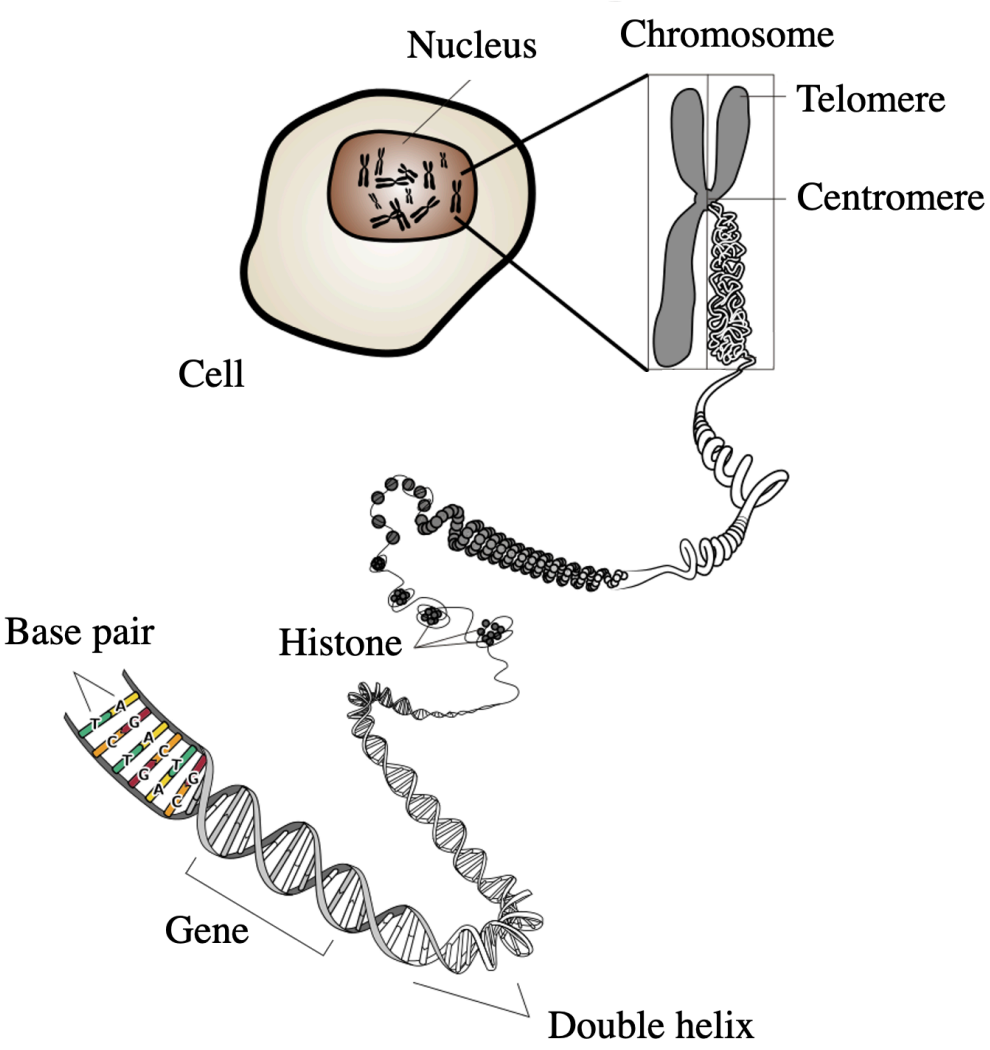

Figure 1.1: Basic concepts of molecular biology. Chromosomes are located in the nucleus of a cell. Each chromosome, essentially a DNA molecule, consists of two polynucleotide chains forming a double helix structure. Two strands are connected by A-T and G-C base pairs. A segment of a DNA codes for a gene. Note: This figure has been derived from a Wikimedia file available in the public domain.

each of which has one of the four types of *nucleobases*, or simply *bases*: adenine (A), cytosine (C), thymine (T), and guanine (G). It is the sequence of these four bases that encodes the genetic information and controls the function of the cells. Notably, adenine (A) in one strand binds with thymine (T) in the other strand forming an A-T base pair. Similarly, cytosine (C) binds with guanine (G) forming a C-G base pair. Thus, the double-stranded DNA molecule is a long chain of base pairs. As one base determines the other base in a base pair, we can represent the DNA as string of A's, C's, T's, and G's – a chain of bases. Human *genome* – the complete genetic code – consists of two copies of DNA molecules from 23 different chromosomes, each copy having a total of over 3 billion base pairs (Ensembl Genome Browser 102, 2020).

1.1.3 Genetic variants

About 99.9% of the 3 billion base pairs of the genome are identical between two human beings. Genetic variation in the rest 0.1% of the genome has the potential to explain inter-individual differences. For example, people with certain variations nearby the *MC1R* gene on chr16 are likely to have red hair (Valverde et al., 1995). Mutation (variation) at a certain genomic position may provide us a clue about an individual's predisposition to a disease. The mutation where a single nucleotide is substituted with another nucleotide is called a *single nucleotide variant (SNV)*. If the SNV is detected in a sufficiently large fraction (e.g., at least 1%) of a population, it is called a *single nucleotide polymorphism (SNP)*. Mutations other than single nucleotide substitution lead to other types of genetic variations e.g., an insertion or deletion of bases (also

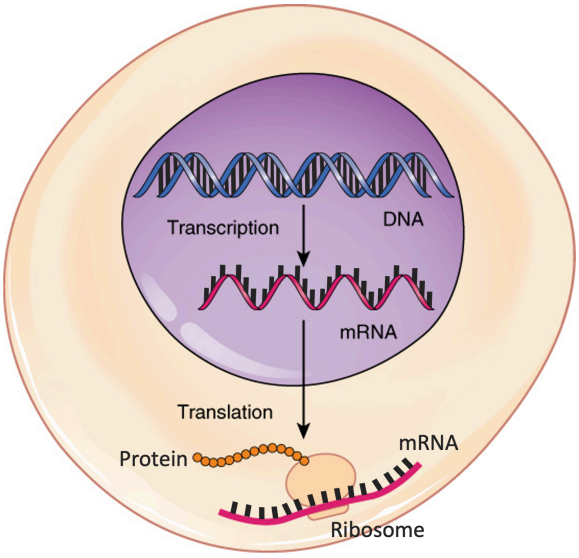

Figure 1.2: The central dogma of molecular biology. DNA is transcribed to RNA which is then translated to protein. Note: this figure has been derived from a Wikimedia file available under CC BY 4.0 license.

known as *indels*), duplication of bases, inversion of a DNA segment, copy-number variation, translocation, etc. SNPs and indels are the most common types of variants in humans. Different forms of a variant (e.g., normal and mutant) are referred to as *alleles*.

1.1.4 The central dogma of molecular biology

The central dogma of molecular biology explains how genetic information in DNA flows in organisms. A segment of a DNA that contains the instructions for a specific molecule, either ribonucleic acid (RNA) or protein, is called a *gene* (see Figure 1.1). According to the central dogma of molecular biology, the double-stranded DNA of a gene is used as a template to create a single-stranded messenger RNA (mRNA). Like DNA, an RNA molecule also has

4 types of bases: adenine (A), cytosine (C), uracil (U), and guanine (G). A thymine (T) in DNA is replaced by an uracil (U) in RNA. The process of creating an RNA molecule from a DNA molecule is called *transcription*. mRNA created from transcription moves from the nucleus to the cytoplasm. There, ribosomes read three RNA bases at a time (codon) to produce protein. The process of creating a protein molecule from an RNA is called *translation*. Proteins are generally considered as the *workhorse of life* determining downstream phenotypes.

The process of producing RNA or protein from a gene is called *gene expression*. Transcription and translation determine gene expression i.e. the amount of RNA or proteins produced from a gene. Because of these processes, even though the genetic information in DNA is largely same in every cell of an individual, different tissues or cell types can perform distinct functions.

Humans have about 20,000 protein-coding genes. Notably, not all genes code for proteins. Genes that do not code for proteins are called noncoding genes. RNAs produced from these genes are called noncoding RNAs. Examples include transfer RNA (tRNA), ribosomal RNA (rRNA), small nuclear RNA (snRNA), micro RNA (miRNA), long noncoding RNA (lncRNA), etc. There are thousands of noncoding genes in humans. They participate in many cellular processes including RNA splicing.

1.1.5 RNA Splicing

The single-stranded RNA synthesized from the DNA template of a gene is called a pre-mRNA (pre messenger RNA). In eukaryotes, including humans,

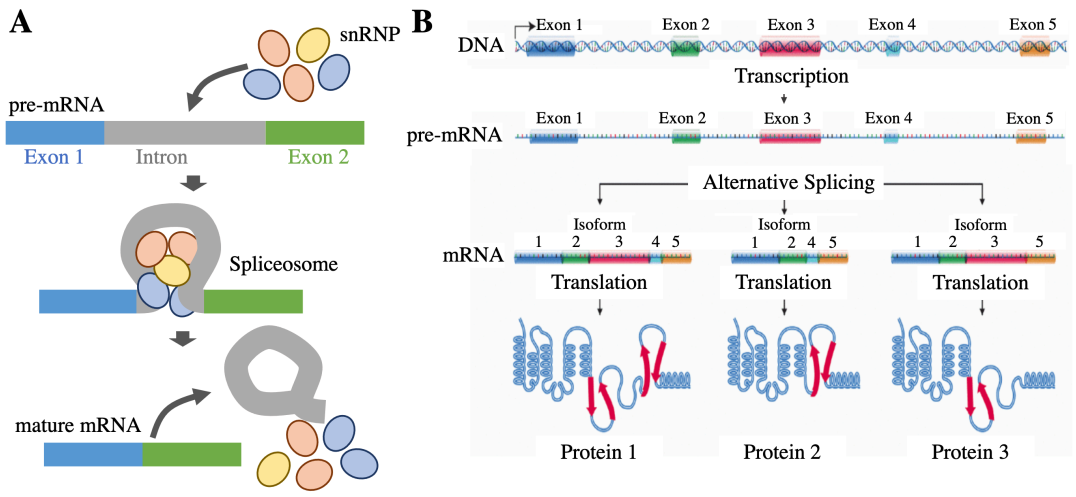

Figure 1.3: RNA Splicing. A) RNA splicing produces mature mRNA from a pre-mRNA. snRNPs and other proteins bind to pre-mRNA to form a spliceosome. Spliceosome brings two ends of an intron together to make a loop. Finally, the intron loop is excised and the neighboring exons are joined together to form a mature mRNA. B) Alternative splicing produces different isoforms from the same pre-mRNA produced by transcription. In this example, each of the three isoforms is translated to a different protein. Note: Figure B has been derived from a Wikimedia file available in the public domain.

pre-mRNAs may be further processed in several steps including polyadenylation, capping, and splicing. *RNA splicing* is an important process where some segments of pre-mRNA, called *introns*, are removed and the remaining segments, called *exons*, are joined together to form a *mature mRNA* or simply *mRNA*. During this process, a number of small ribonucleoproteins or snRNPs (RNA-protein complex) and other proteins form a *spliceosome* that brings two ends of an intron together making a loop. The intron loop along with snRNPs detaches, and the ends of neighboring exons join together (see Figure 1.3A).

In *alternative splicing*, exons of the pre-mRNA are connected in different ways to produce different mature mRNAs or *isoforms* (see Figure 1.3B). Due to the difference in the sequence, each isoform may produce a different protein. Thus, alternative splicing enables producing multiple proteins from the same gene. Each protein isoform may have a different function, therefore, alternative splicing has the potential to control the downstream phenotype. In fact, alternative splicing events are up to 30% more common in tumors than normal samples (Kahles et al., 2018).

1.1.6 Regulation of gene expression

Genes encode the instructions to produce proteins and proteins perform cell functions. The type and the amount of proteins present in a cell critically defines the function of the cell and consequently the function of an organism. Disruption of this delicate balance of the type and the amount of proteins present in a cell often leads to clinically relevant phenotypes. Thus, it is

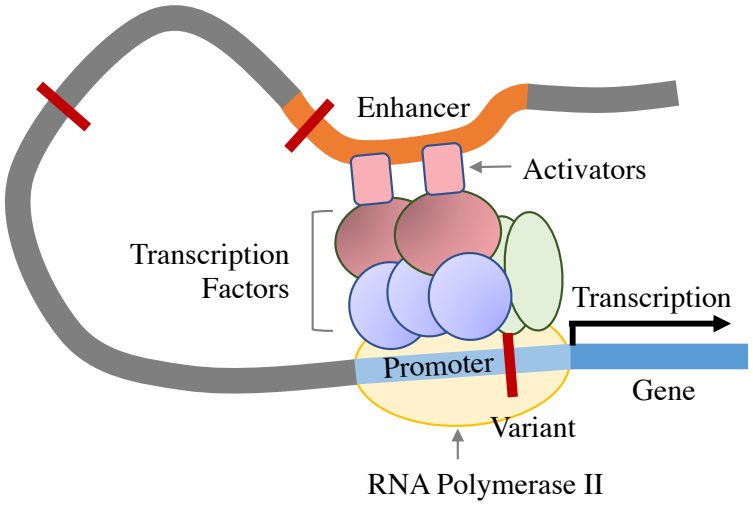

Figure 1.4: Schematic of transcription. Transcription factors bind to the promoter region to recruit RNA polymerase II that starts transcription. Activators/Repressors and genetic variants may control the transcription process.

crucial to understand how this balance is maintained or disrupted in disease. Importantly, gene expression regulation is affected by both genetic and environmental factors.

Transcription is a fundamental way to control gene expression. *Transcription factors (TFs)* play an important role in the transcription of a gene. TFs bind to the promoter region nearby the transcription start site (TSS) of a gene to recruit RNA polymerase II which transcribes the gene (see Figure 1.4). Binding of activators in the enhancer regions (relatively distant from the TSS) may increase the likelihood of transcription. Binding of repressors to the DNA may prevent transcription. Thus, the transcription of a gene can be modulated by any genes affecting the transcription machinery. Notably, the transcription machinery can also be affected by presence or absence of genetic variants. For

example, the presence of a certain variant may disrupt the transcription factor binding site and consequently prevent transcription.

Splicing is another major contributor to control gene expression. The splicosome is dynamically comprised of hundreds of proteins. Splicing regulatory elements recruit sequence-specific RNA-binding protein factors, known as splicing factors, that either activate or repress splice site recognition or spliceosome assembly.

Gene expression could also be regulated epigenetically by controlling access to the transcription machinery. DNA methylation and histone modifications are some of the common ways to regulate DNA accessibility and chromatin structure.

1.1.7 DNA and RNA sequencing

DNA sequencing is the process of determining the correct order of nucleic acids in DNA. Though the first genome was sequenced in the 1970s, genome sequencing became cost-effective relatively recently because of high-throughput next-generation sequencing (NGS). NGS breaks up the DNA randomly into a large number of small fragments, sequences them in parallel, and finally stitches the small sequences together using computational algorithms to reconstruct the original genome (Langmead & Salzberg, 2012; Langmead et al., 2009; Li & Durbin, 2009).

NGS can sequence RNAs as well by reverse-transcribing RNAs to complementary DNAs (cDNAs) followed by sequencing (Figure 1.5). Given NGS

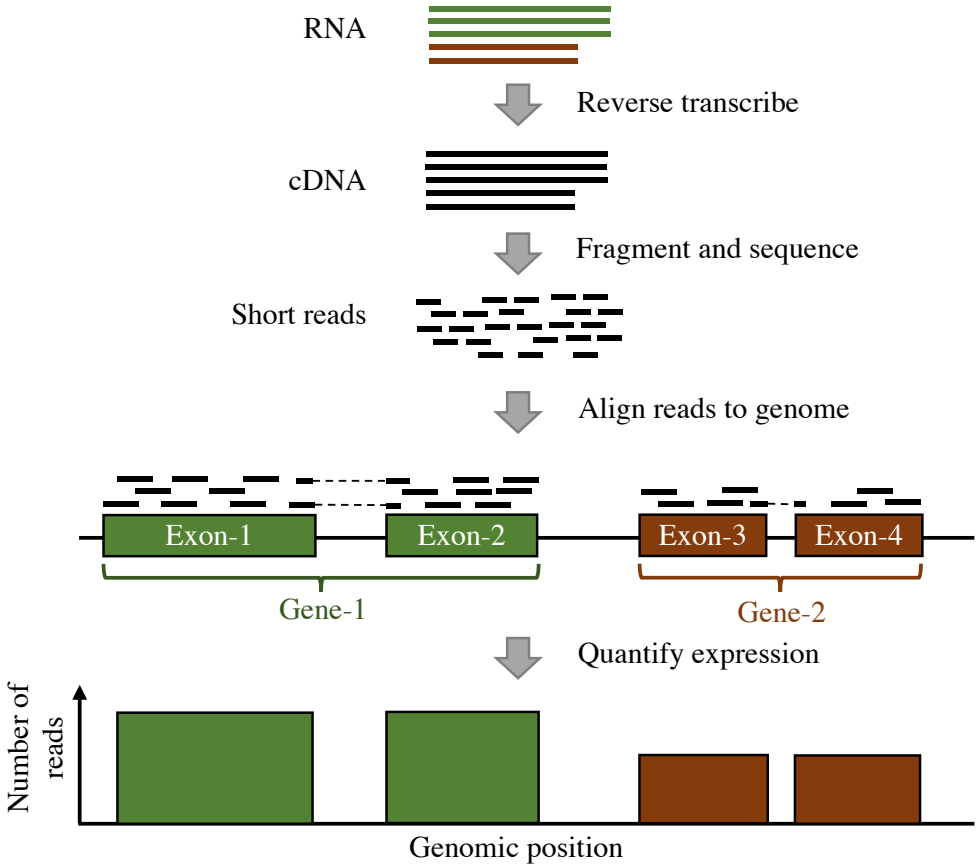

Figure 1.5: Schematic of sequence alignment and expression quantification in RNA sequencing. After collection and preparation of RNA samples, reverse transcription produces cDNA. Fragmentation of cDNA followed by sequencing produces short reads. Short reads are aligned to the reference genome and expression is quantified.

sequences only short reads, two steps are critical in RNA sequencing (RNA-seq): sequence alignment and expression quantification. During sequence alignment, each read is mapped (or aligned) to a reference genome to determine which genomic region was transcribed to form the corresponding RNA. Sequence alignment algorithms look for the degree of similarity between each read and the reference genome to find the correct region. Splicing-aware aligners are good candidates for aligning RNA-seq reads (Dobin et al., 2013; Kim et al., 2015; Kim et al., 2019; Kim et al., 2013; Trapnell et al., 2009).

Expression quantification is the process to estimate the abundance of reads aligned to each gene, isoform, exon, or other levels. While the simplest approach is to count the number of reads aligned to a genomic region, special attention may be required to disambiguate overlapped regions (Anders & Huber, 2010; Kovaka et al., 2019; Li & Dewey, 2011; Love et al., 2014; Patro et al., 2017; Pertea et al., 2015). Notably, some software tools can quantify expression without alignment making the process fast (Bray et al., 2016; Patro et al., 2017).

Recent technologies can sequence long reads though the error rate is slightly higher than short reads (Chin et al., 2013; Greninger et al., 2015). Currently, single-cell sequencing is gaining popularity (Macosko et al., 2015; Picelli et al., 2014; Xia et al., 2019; Zheng et al., 2017).

1.2 Computational frameworks

In this thesis, we use DNA and RNA sequencing data to decipher gene regulatory programs in humans. We study how expression of a gene is regulated by

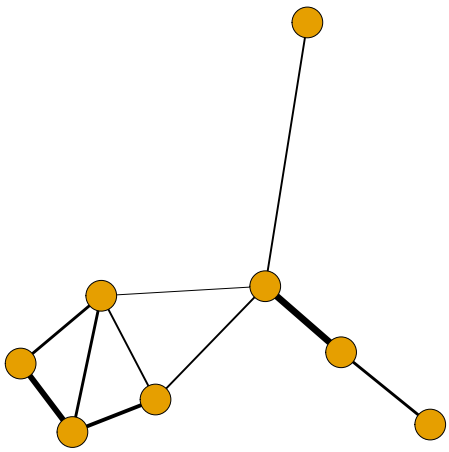

Figure 1.6: Gene co-expression network (GCN). Each node represents a gene and each edge connecting two nodes represents co-expressed gene pairs. The width of each edge is proportional to their co-expression.

other genes and genetic variants following two computational frameworks, co-expression networks and quantitative trait loci (QTL), respectively.

1.2.1 Co-expression networks

A gene co-expression network (GCN) is schematically represented by an undirected graph where each node represents a gene and a pair of nodes are connected by an edge if they are significantly co-expressed (Stuart et al., 2003) (Figure 1.6). Each edge may have a weight, giving rise to a weighted gene co-expression network, where the weight represents the strength of co-expression. A weight of zero means that the nodes are not co-expressed. Connected genes generally have similar functions or they are involved in a common biological

process.

The structure of a gene co-expression network is an abstraction of the molecular dynamics in a particular context. It represents biological pathways – the cascading of information flow in a biological system. A gene co-expression network can help us understand biological processes. Applied in a disease context, it can help us discover disease mechanisms and find appropriate drug targets (Mei et al., 2012; Saha et al., 2015; Saha et al., 2014).

WGCNA, Weighted Gene Co-expression Network Analysis (Langfelder & Horvath, 2008; Zhang & Horvath, 2005), is one of the simplest and most popular methods to infer a co-expression network from gene expression data. It estimates the edge weight (W_{ij}) between a pair of genes as their co-expression similarity raised to a power β.

$$W_{ij} = cor(x_i, x_j)^{\beta} \tag{1.1}$$

Here, x_i and x_j represent expressions of gene i and gene j, respectively, and $\beta \geq 1$ is selected in such a way that the network follows a scale-free topology. Though simple in nature, WGCNA successfully found modules in many different contexts (DiLeo et al., 2011; Hartl et al., 2020; Yin et al., 2018; Zhai et al., 2017). A limitation of WGCNA is that it cannot distinguish between direct edges and indirect edges via other genes.

Graphical lasso (Friedman et al., 2008), another popular network inference method, attempts to learn potentially direct edges between genes by learning a precision matrix (also known as an inverse covariance matrix, $\hat{\Theta}$) with L-1

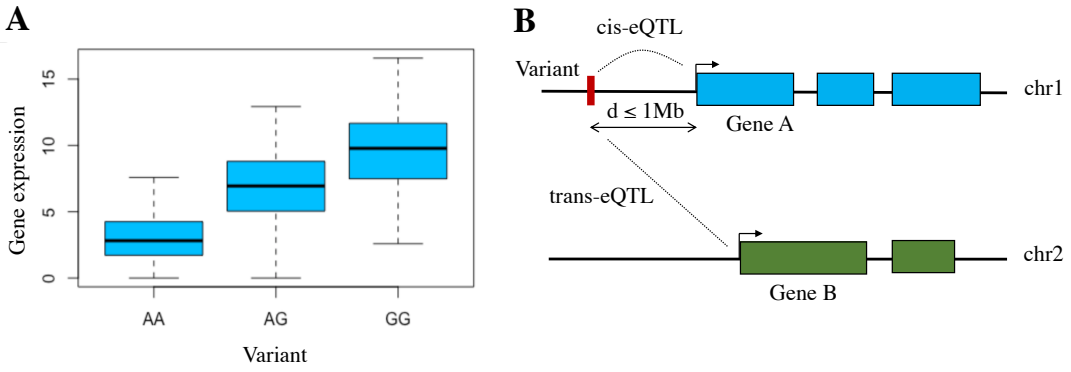

Figure 1.7: Expression quantitative trait loci (eQTL). A) A variant on the x-axis explains the expression of a gene on the y-axis to form an eQTL (variant-gene pair). B) Cis- and trans-eQTLs.

regularization.

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} - \log\det\ \Theta + tr(S\Theta) + \lambda\|\Theta\|_1 \tag{1.2}$$

Here, S is the sample covariance matrix and λ is the penalty parameter. A nice property of the precision matrix is that a non-zero value represents conditional dependency between the corresponding genes given other genes. However, graphical lasso is very slow for a high number of genes, and often the selection of an optimal lambda using cross-validation is practically infeasible.

Examples of other popular co-expression network inference methods include ARACNE (Margolin et al., 2006), CLR (Faith et al., 2007), MRNET (Meyer et al., 2007), MRNETB (Meyer et al., 2010), and GENIE3 (Huynh-Thu et al., 2010).

1.2.2 Quantitative trait loci (QTLs)

A quantitative trait loci (QTL) study finds genetic variants explaining the variation in a quantitative trait (Nica & Dermitzakis, 2013). When the trait is the expression of a gene, it is called an expression quantitative trait loci (eQTL) study. Figure 1.7A illustrates that the variant on the x-axis determines the level of expression of a gene on the y-axis. Formally, we test for a linear association between gene expression y and genotype x.

$$y = \alpha + \beta x + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{1.3}$$

When x and y are significantly associated ($\beta \neq 0$), we call the variant-gene pair an *expression quantitative trait locus (eQTL)*. The variant and the gene are referred as an *eVariant* and an *eGene*, respectively. β is generally referred as the *effect size*.

There are two types of eQTLs: *cis* and *trans*. A cis-eQTL acts on the same DNA molecule, and a trans-eQTL acts on a different molecule. We approximate cis and trans-eQTLs based on the distance between the variant and the gene. If the variant resides nearby the gene, typically within 1 Mb of the gene's transcription start site (TSS), the variant-gene pair is called a *cis-eQTL* (Figure 1.7B). In contrast, if the variant resides away from the gene, typically on a different chromosome, the variant-gene pair is called a *trans-eQTL*. Finding trans-eQTLs is more challenging than cis-eQTLs because of the high number of tests and relatively small effect sizes. Any technical bias can further complicate the problem. We focus on trans-eQTLs in this thesis.

In both cis- and trans-eQTLs, the variant likely regulates the expression of the gene, or the variant is in linkage disequilibrium (LD) with the variant regulating the gene. EQTLs provide a mechanism to explain how a genetic variant may control a downstream phenotype through regulating intermediate molecular phenotypes such as gene expressions (Dutta et al., 2020; Gill et al., 2020; Hawe et al., 2020; The GTEx Consortium, 2017, 2020; Võsa et al., 2018).

1.3 Challenges

A common challenge in computation genomics is the limited number of samples. While we can quantify more than 20,000 genes with the advent of recent sequencing technologies, the number of samples is still in the range of hundreds for a specific tissue or study. The challenge gets further complicated for co-expression networks and trans-eQTLs because of millions of parameters estimation in co-expression networks and billions of tests in trans-eQTLs.

Any systematic error in sequencing experiments, sequencing alignment, expression quantification, or any other data processing pipeline may produce spurious associations in the data affecting both co-expression networks and trans-QTL studies. If not properly handled, batch effects and unmeasured technical factors may confound both types of studies. Even random white noise in data could be critical because of the relatively small sample size compared to the number of tests.

A big challenge in a co-expression network estimation is to distinguish between direct and indirect edges. The presence of collinear genes makes the challenge harder. Complex models easily overfit the data because of a

relatively small number of samples. Advanced models might be prohibitively slow to run. Besides, appropriate biological ground truths generally do not exist, making the network evaluation difficult.

A trans-eQTL analysis takes a long time and memory to run. It also takes a large amount of disk space to save the eQTL statistics. Besides, multiple testing correction is tricky for a trans-eQTL analysis because of linkage disequilibrium between genetic variants.

1.4 Thesis outline

We address some of the above challenges in the remaining chapters.

- In Chapter 2, we investigate a potential source of technical errors in trans-eQTL and co-expression network studies. We study the prevalence of such technical errors in a standard pipeline using human data. We also present an approach to detect and avoid potential false positives due to these errors. This chapter is based on a published work from Saha and Battle, 2018.

- In Chapter 3, we turn to a network system to jointly study the regulation of transcription and alternative splicing. By appropriate modeling of RNA-sequencing data, we attempt to find potential splicing regulators in humans. We further analyze tissue-specificity of transcription and splicing regulatory relationships and detect genetic variants supporting these relationships. This chapter is based on a published work from Saha et al., 2017.

- In Chapter 4, we present a novel network inference method named *SPICE* to prioritize potential direct regulatory relationships in transcription. We also present a comprehensive set of metrics to evaluate networks using biological data.

- Finally, in Chapter 5, we conclude the thesis by summarizing the findings and discussing future directions.

Chapter 2

Cross-mappability: False positives in trans-eQTL and co-expression studies

Expression quantitative trait loci (eQTL) and co-expression networks, as we described in Chapter 1, are two effective and popular tools to study gene regulation. These methods generally make a large number of tests and/or estimate a large number of parameters using a limited number of data samples. Consequently, these methods have limited power to detect true regulatory relationships. Any systematic error in the pipeline from sample collection to statistical analysis, if not handled properly, can result into false positive discoveries. It is critical to detect the sources of false discoveries and take necessary steps to avoid related consequences. In this work, led by me, we assessed the potential for incorrect alignment of RNA-sequencing reads to cause false positives in both gene expression quantitative trait loci (eQTL) and co-expression analyses. Our main contributions are as follows:

- Trans-eQTLs identified from human RNA-sequencing studies appeared

to be particularly affected by alignment errors due to sequence similarity, even when only uniquely aligned reads are considered. Over 75% of trans-eQTLs detected using a standard pipeline occurred between regions of sequence similarity and therefore could be due to alignment errors.

- Associations due to mapping errors are likely to misleadingly replicate between studies.

- To help address this problem, we quantified the potential for *cross-mapping* to occur between every pair of annotated genes in the human genome. Such cross-mapping data can be used to filter or flag potential false positives in both trans-eQTL and co-expression analyses.

This work was published in F1000 Research (Saha & Battle, 2018), and this chapter is based on the published article.

2.1 Introduction

Sequence similarity among distinct genomic regions makes alignment of short sequencing reads difficult (Johnson et al., 2016; Kahles et al., 2016). Genomes, including the human genome, contain diverse classes of elements with sequence similarity across regions, ranging from large segmental duplications to pseudogenes to transposable elements. Alignment-based quantification of genomic phenotypes such as gene expression or epigenetic signal is less reliable for such regions (Degner et al., 2009; Derrien et al., 2012; Karimzadeh et al., 2018; Robert & Watson, 2015).

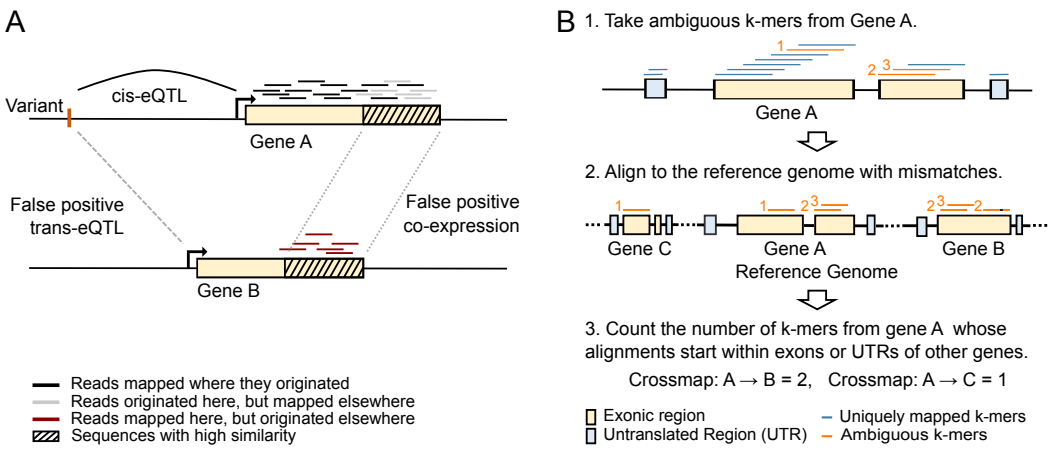

Figure 2.1: Overview of cross-mappability. A) Some of the reads generated from Gene A are incorrectly mapped to Gene B because of sequence similarity between the genes, leading to false positive co-expression. Consequently, a variant which is a true cis-eQTL of Gene A appears as a false positive trans-eQTL of Gene B. B) We align the ambiguous (orange) *k*-mers (75-mers from exons and 36-mers from UTRs) from Gene A to the reference genome using Bowtie and count how many *k*-mers from Gene A map to each other gene to compute cross-mappability. Here, the number beside each ambiguous (orange) *k*-mer represents the identifier for the ambiguous *k*-mer based on its position in Gene A.

Despite attention to the importance of alignment errors, the full range of consequences is not always considered in downstream analyses. Here, we focus on evidence that sequence similarity between pairs of genes and resulting alignment errors between them may lead to false positives in association studies from RNA-sequencing (RNA-seq) data, specifically in expression quantitative trait locus (eQTL) and co-expression analyses. eQTL studies, revealing associations between genetic variants and gene expression levels, have contributed to a greater understanding of gene regulation and genetics of complex traits (Albert & Kruglyak, 2015; Grundberg et al., 2012; Nica & Dermitzakis, 2013). Trans-eQTLs, where the genetic variant is distant or on a different chromosome from the associated gene, are of particular interest, but have proven challenging to identify in human data due to power, confounders, small effect sizes, and other challenges (The GTEx Consortium, 2017; Westra et al., 2013). Given that a trans-eQTL analysis performs genome-wide tests, it is more prone to be affected by systematic errors between genomic regions than a cis-eQTL analysis where only variants close to the target gene are considered. Here, we discuss the impact of alignment errors on RNA-seq association studies. Figure 2.1A illustrates a cartoon example, where all reads truly originate from transcripts of Gene A, but due to sequence similarity between Gene A and Gene B, some of the reads incorrectly map to Gene B, causing it to erroneously appear to be expressed in the sample. The number of reads misaligned to Gene B across samples may be directly proportional to the number of reads for Gene A, or may be determined by genetic variation creating sequence mismatches with the correct region. In either case, spurious associations can then arise. In Figure 2.1A, the two genes incorrectly appear to

be co-expressed. In addition, a variant associated with expression of Gene A may also appear to be associated with Gene B, giving rise of a false positive trans-eQTL. We note that such errors are not entirely mitigated by filtering multi-mapped reads—some alignment errors may remain between similar regions even among uniquely aligned reads due to genetic variation, errors in the reference genome, and other complications.

Previous studies have shown that uniqueness of sequence in genomic regions should be considered in an analysis of sequencing data (Derrien et al., 2012; Karimzadeh et al., 2018; van de Geijn et al., 2015). Karimzadeh et al. showed that a differential methylation analysis can identify false signals due to poor mappability (Karimzadeh et al., 2018). We have previously filtered trans-eQTLs based on sequence similarity as part of the Genotype-Tissue Expression (GTEx) project (The GTEx Consortium, 2017) and the Depression Genes and Networks (DGN) study (Battle et al., 2014). Pickrell et al. (Pickrell et al., 2010) also suggested that the most significant distant eQTL in their RNA-seq study was likely an artifact arising due to sequencing reads originating from a gene near the SNP mapping to another distant gene. Related effects were also discussed in greater depth for microarrays, where probes intended for one gene may cross-hybridize to other genes (Reilly et al., 2006; Westra et al., 2013). In microarray studies, one could identify and replace probes displaying poor specificity, but in RNA-seq, any region of sequence similarity between genes can cause alignment errors. Previous studies have not presented a systematic analysis of alignment-related false positives in RNA-seq association testing.

Here, we report the prevalence of potential false positives in trans-eQTL

and co-expression analyses arising from alignment errors. We present a method to assess the potential for mapping error between pairs of genes, which can then be used to filter or flag associations that could arise from these errors. We introduce a new metric, *cross-mappability*, representing the extent to which reads from one gene may be mapped to another gene. Using gene expression data from GTEx (The GTEx Consortium, 2017) and DGN (Battle et al., 2014), we demonstrate the impact of misalignment on both trans-eQTL detection and co-expression analysis in real data. Notably, we show that over 75% of trans-eQTLs detected in any GTEx tissue using a naive pipeline are potential false positives, emphasizing that it is critical to consider these errors. To support future studies, we have published codes in Github (Saha & Battle, 2019a) and also made cross-mappability resources publicly available for the human genome (hg19 and GRCh38) (Saha & Battle, 2019b).

2.2 Methods

2.2.1 Mappability and cross-mappability

We developed a new metric, *cross-mappability*, to quantify the potential for incorrect read alignment where reads originating from one gene may incorrectly map to another gene. Based on annotated transcripts for each gene, we evaluated k-mers from exonic and untranslated regions (UTRs) of the reference genome that serve as a proxy for reads in an RNA-seq experiment. We defined cross-mappability from Gene A to Gene B, *crossmap(A, B)*, as the number of Gene A's k-mers whose alignment, allowing mismatches, start within exonic or untranslated regions of Gene B. Notably, existing *mappability*

scores (Derrien et al., 2012; Karimzadeh et al., 2018) correspond to a single region (or gene) describing uniqueness of the sequence of the region in the genome, our *cross-mappability* score corresponds to a pair of genes describing similarity between the sequences of the genes.

Though cross-mappability is a straightforward metric, its computation is non-trivial due to the size of the genome. We followed a systematic approach to compute genome-wide cross-mappabilities in practice. Following Derrien et al., 2012, we define mappability of a k-mer as $\frac{1}{C_k}$, where C_k is the number of positions where the k-mer maps to the genome with a tolerance of up to 2 mismatches. We computed exon- and UTR-mappability of a gene as the average mappability of all k-mers in exonic regions and untranslated regions, respectively. We used a collapsed gene model to generate k-mers where overlapped exons and overlapped UTRs were merged to form exonic and UTR regions, respectively. Then, mappability of a gene is computed as the weighted average of its exon- and UTR-mappability, weights being proportional to the total length of exonic regions and UTRs, respectively. Importantly, we only have to compute cross-mappability from genes with mappability < 1, as no k-mer from a gene with mappability $= 1$ will map to other regions of the genome (i.e. these will all result in cross-mappability of 0). Moreover, we need to consider only k-mers with mappability < 1 from a gene, as uniquely mapped k-mers will not map to other genes. So, we align all such k-mers from exonic and untranslated regions of each gene to the reference genome using Bowtie v1.2.2 (Langmead et al., 2009), tolerating up to 2 mismatches, and then count the number of k-mers whose alignment

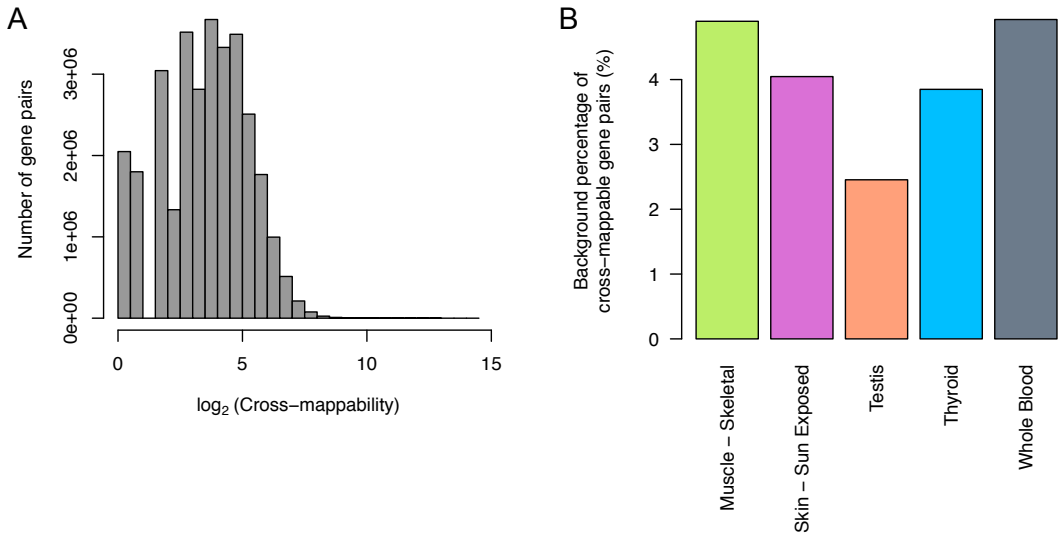

Figure 2.2: Cross-mappability statistics. Cross-mappability statistics. A) Distribution of cross-mappability between cross-mappable pairs of genes, restricted to gene pairs with cross-mappability > 0, using Gencode v19 annotations on human genome hg19. B) Background percentage of cross-mappable gene pairs between all available expressed genes in GTEx data, categorized by tissue. For both panels, directed gene pairs were used; i.e., (Gene A, Gene B) and (Gene B, Gene A) pairs were considered different.

start within exonic or untranslated regions of every other gene to compute cross-mappability with each gene genome-wide (Figure 2.1B).

The length k may be tuned to match particular read length or alignment method. Here, if the value of k is not mentioned for k-mers, the default value of k is 75 for exons and 36 for UTRs. We used a smaller k for UTRs than for exons because UTRs are generally shorter than exons. Mappability of a gene and cross-mappability to/from a gene is undetermined if all the exons of the gene are shorter than 75 bp and all the UTRs are shorter than 36 bp.

We computed genome-wide mappability and cross-mappability for human genome hg19 using annotations from Gencode v19 (Harrow et al., 2012).

26,200 (out of 57,820) genes had at least one *k*-mer cross-mapping to/from another gene. There were 31,167,448 gene pairs (0.93%) that were cross-mappable (cross-mappability > 0). Figure 2.2A shows the cross-mappability distribution. We found that 2.45-4.92% of gene pairs expressed and quantified in five tissues of the GTEx v7 data were cross-mappable (Figure 2.2B). We also computed the same set of resources for human genome GRCh38 using annotations from Gencode v26, all of which are publicly available (Saha & Battle, 2019b).

2.2.2 Data

We downloaded fully processed, filtered and normalized gene expression data used in GTEx eQTL analysis from the GTEx portal (www.gtexportal.org). For this study, we focused on gene expression data from 5 tissues: whole blood, skeletal muscle, thyroid, sun-exposed skin, and testis. We also obtained covariates including 3 genotype PCs representing ancestry, sex, genotyping platform, and PEER factors (Stegle et al., 2012) as released in GTEx v7. GTEx aligned 76-bp paired-end reads to the reference genome with STAR v2.4.2a (Dobin et al., 2013), quantified gene expression levels with RNA-SeQC v1.1.8 (DeLuca et al., 2012) using uniquely mapped reads aligned in proper pairs and fully contained within exon boundaries where each alignment must not contain more than six non-reference bases. We downloaded genotype data from GTEx release v7 from dbGaP (accession number: phs000424.v7.p2).

We also collected genotype, processed RNA-seq, and covariate data for the DGN cohort, which is available through the National Institute of Mental

Health (NIMH) Center for Collaborative Genomic Studies on Mental Disorders. DGN aligned the reads to the reference genome using TopHat (Trapnell et al., 2009) and quantified gene expression levels using HTSeq (Anders et al., 2015). Latent factors inferred from the expression data have already been regressed out of the processed DGN data to address hidden confounders, as described by Battle et al., 2014. Gene symbols were mapped to Ensembl gene ids using Gencode v19.

We downloaded the list of trans-eQTLs in 33 cancer types detected by PancanQTL (Gong et al., 2018) from http://bioinfo.life.hust.edu.cn/PancanQTL. For consistency with our study, we used trans-eQTLs where the variant and the gene were on different chromosomes, and the gene symbols were mapped to unique Ensembl gene ids according to Gencode v19.

2.2.3 Trans-eQTL detection

For trans-eQTL analysis, we selected autosomal variants with MAF ≥ 0.05 that did not fall in a repeat region as annotated by the UCSC RepeatMasker track (Casper et al., 2017). We tested trans-eQTL association for each inter-chromosomal variant-gene pair using Matrix-eQTL's linear model test (Shabalin, 2012). For GTEx, three genotype PCs, genotyping platform, sex, and PEER covariates estimated by GTEx were used as covariates in Matrix-eQTL. We computed the false discovery rate using the Benjamini-Hochberg method within each tissue. The covariates used for trans-eQTL replication in DGN were three genotype PCs, sex and age, as the expression data already had latent factors regressed out.

2.2.4 Co-expression analysis

We quantified co-expression of a pair of genes as the absolute Pearson correlation ($|r|$) between expression levels of the genes across all available samples. For GTEx, we regressed out all covariates including PEER factors before co-expression analysis. For DGN, we used the corrected data which also regresses out latent factors.

2.3 Results

2.3.1 Effect of cross-mappability on trans-eQTL detection

To investigate the effects of alignment errors on trans-eQTL detection, we performed a standard trans-eQTL analysis using data from the GTEx project for five human tissues. For this study, we categorized an eQTL as "cis" if the variant is within 1Mb of the transcription start site (TSS) of the gene, and "trans" if they are on different chromosomes, approximating the regions where cis and trans mechanisms are likely to occur. We call a trans-eQTL "cross-mappable" if any gene within 1Mb of the identified trans-eQTL variant cross-maps to the trans-eQTL target gene. The cross-mappable trans-eQTLs represent suspicious hits that could potentially arise simply due to alignment errors, although cross-mappability does not definitively establish that any individual trans-eQTL is a false positive.

We identified 19,348 unique trans-eQTLs (variant-gene pairs) at FDR $\leq$ 0.05 from five tissues corresponding to 14,785 unique SNPs and 1,419 unique

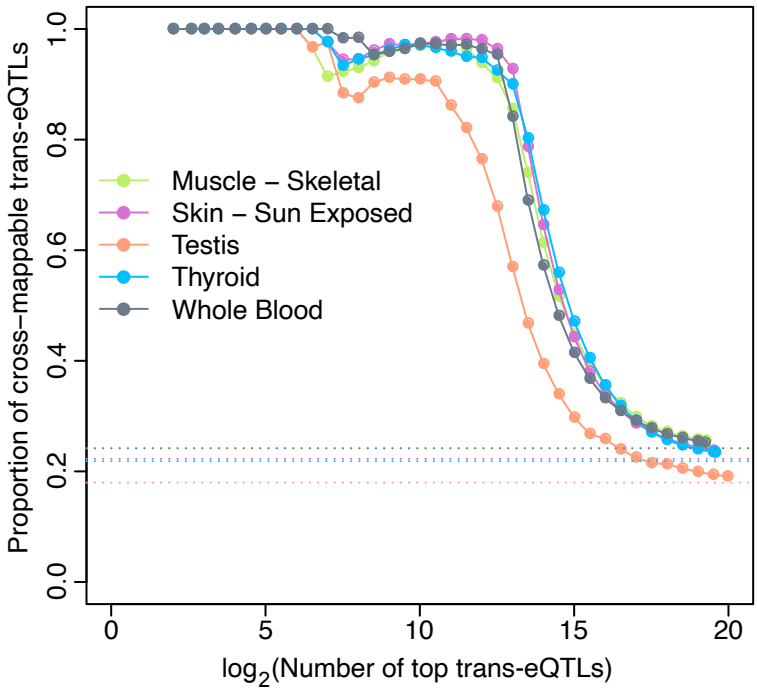

Figure 2.3: Effect of cross-mappability on trans-eQTLs in GTEx. Fraction of cross-mappable trans-eQTLs among the top significant variant-gene pairs (ordered by increasing FDR) in each tissue (color). Each dotted horizontal line represents the background cross-mappable rate in a given tissue.

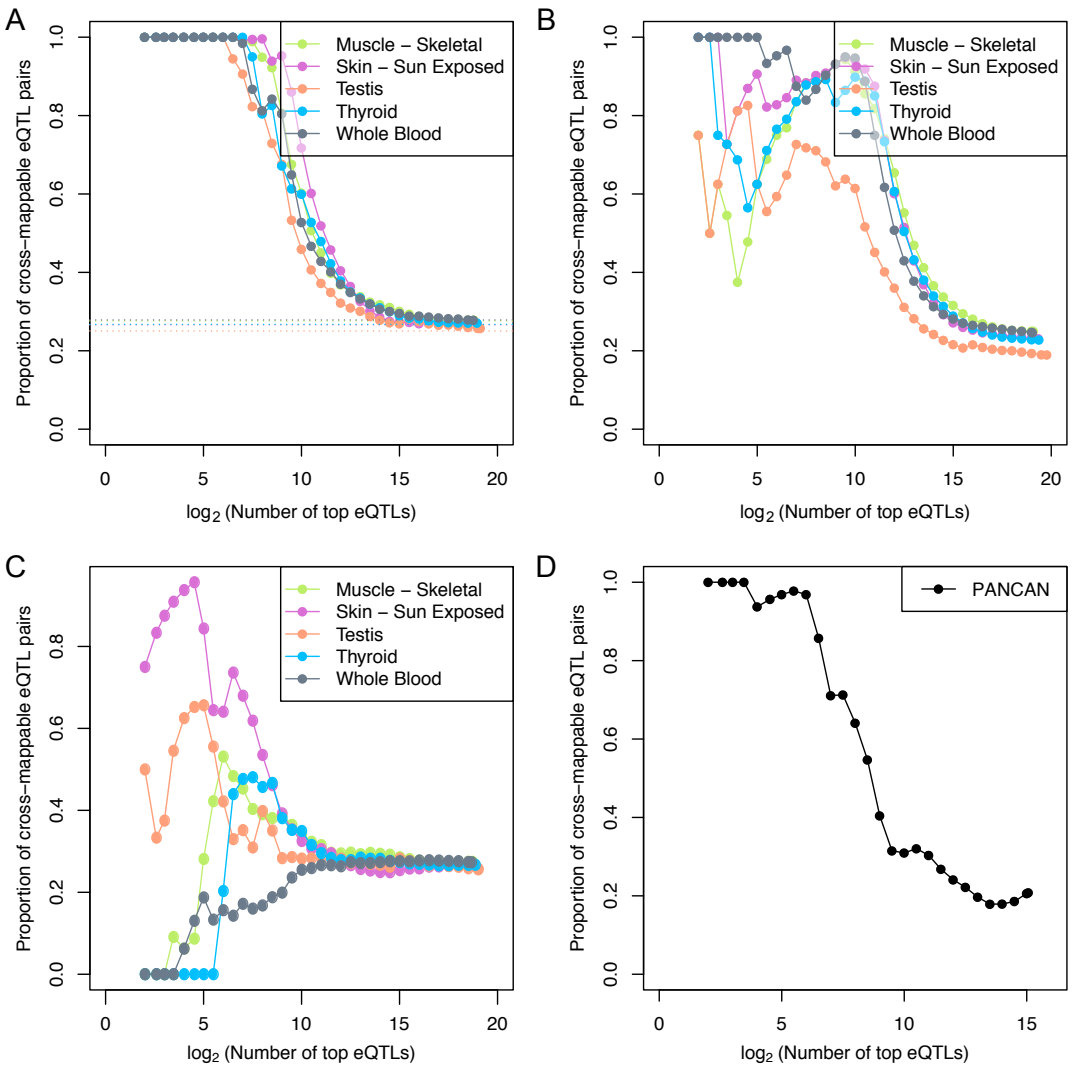

Figure 2.4: Cross-mappability in top trans-eQTL hits A) using protein-coding genes in GTEx, B) using genes with mappability ≥ 0.8 in GTEx, C) using protein-coding genes with mappability ≥ 0.8 in GTEx, and D) by PancanQTL where unique eQTLs were ordered by lowest p-value across all cancer types.

genes. Notably, a large majority (75.14%) of these statistically significant trans-eQTLs were cross-mappable. Furthermore, the cross-mappable eQTLs tended to be the most highly significant (ordered by increasing p-value, Figure 2.3). In GTEx tissues, 90.8-97.3% of top 1000 trans-eQTLs were cross-mappable, compared to a background rate of 19.1-25.6% (based on all tested variant-gene pairs). The fraction of cross-mappable trans-eQTLs is very high even when we restrict our analysis to protein-coding genes or to genes with mappability ≥ 0.8 (Figure 2.4A-C).

We observed a similar pattern in the trans-eQTLs reported from RNA-seq data of 33 cancer types (Gong et al., 2018) (Figure 2.4D). We also observed that randomly selected variant-gene pairs susceptible to cross-mapping yield more trans-eQTLs than randomly selected pairs with no cross-mapping potential (Figure 2.5). Overall, the high fraction of cross-mappable eQTLs among the top associations in multiple tissues and multiple datasets indicates that alignment errors could be a major source of artifacts, dominating legitimate trans-eQTLs. It is also important to note that filtering such prevalent potential false-positives necessitates re-assessing FDR. For example, while 4,809 trans-eQTLs with no evidence of cross-mapping (corresponding to 969 unique genes) were among the 19,348 hits from the original scan of GTEx, only 2,456 (corresponding to 228 unique genes) would appear significant if FDR were reassessed after filtering cross-mapping hits.

When we further analyzed the composition of the 19,348 significant naive trans-eQTLs, we observed a majority (>70%) of cross-mappable eQTLs corresponded to pseudogene targets. The non-cross-mappable eQTLs contained far

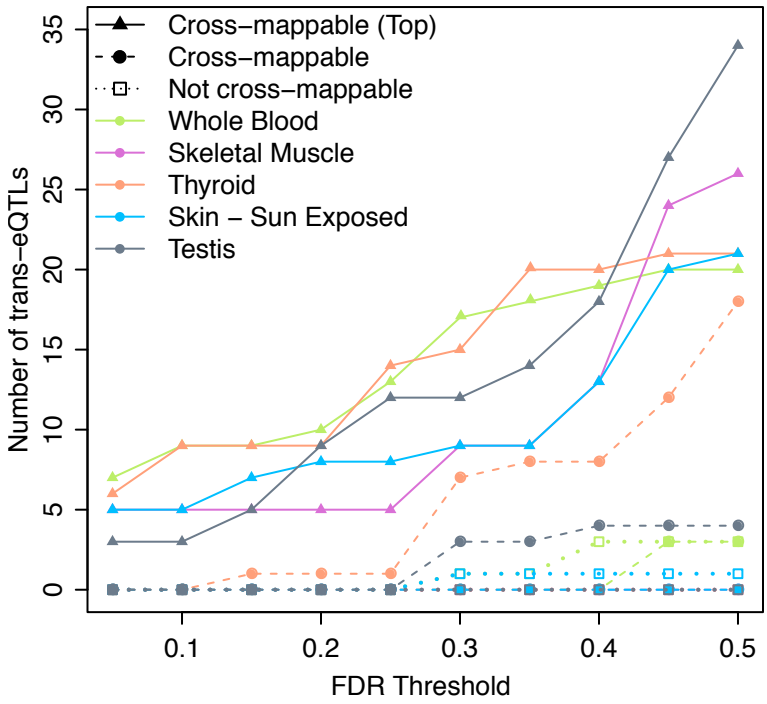

Figure 2.5: Large number of trans-eQTLs among random cross-mappable gene pairs. We tested for trans-eQTLs taking the same number of random variant-gene pairs in 3 different categories: 1) Not cross-mappable, 2) Cross-mappable, and 3) Cross-mappable (Top). In the first category, we randomly selected 1,000 not cross-mappable gene pairs (g1, g2) where g1 and g2 were on different chromosome and there was at least one variant near g1 (within 1Mb of the TSS of g1), then selected the best cis-variant s (with lowest p-value) for g1, and finally tested for trans-association between s and g2. Variant-gene pairs for other two categories were selected in a similar way as the first category except that the gene pairs were cross-mappable (crossmap(g1, g2) > 0) in the second category, and highly cross-mappable (among top 10,000 cross-mappable pairs) in the third category. The above plot shows the number of significant trans-eQTLs (y-axis) detected at a given FDR (x-axis) in each category (line marker) in each tissue (color).

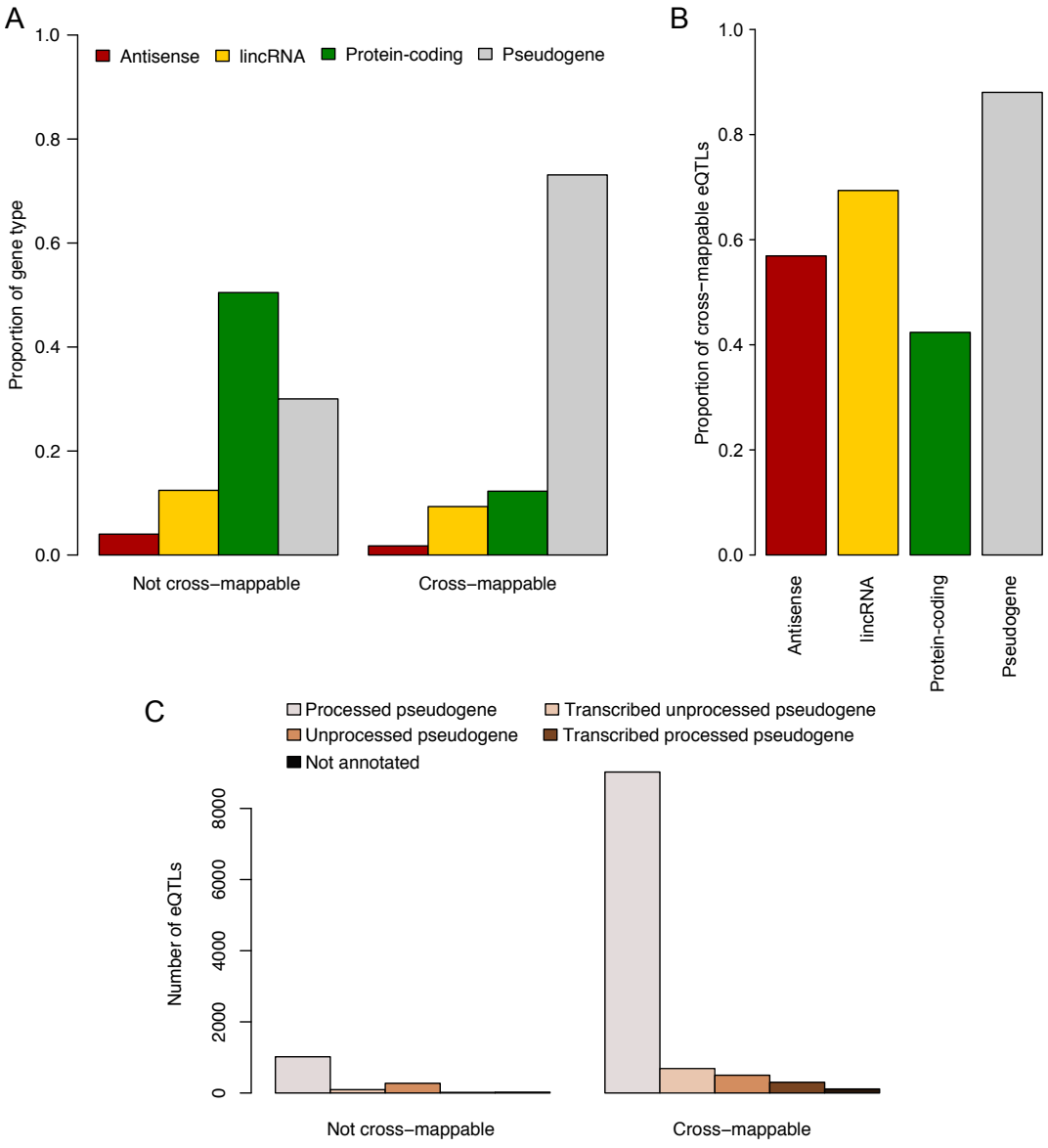

Figure 2.6: Composition of trans-eQTLs. A) Representation of gene types among trans-eQTL target genes, categorized by cross-mappability. B) Proportion of cross-mappable eQTLs categorized by gene type. Only the four most frequent gene types in trans-eQTL hits are shown. C) Among trans-eQTLs with a pseudogene target gene, quantification of different pseudogene sub-types, categorized by cross-mappability. Pseudogene sub-types were identified from the Gencode v26 annotation, as sub-types are not available in Gencode v19. The five most frequent types among trans-eQTL hits are shown.

fewer pseudogene targets (30%, Figure 2.6). Likewise, we observed that more than 85% of eQTLs corresponding to pseudogenes were cross-mappable. Due to sequence similarity between pseudogenes and their corresponding parent genes, this is not surprising and could be due to alignment errors. One simple preventative measure in trans-eQTL studies would be to simply exclude pseudogenes entirely. However, 42.4% of eQTLs corresponding to protein-coding genes were also cross-mappable, which still exceeded expectation, and the top hits remained enriched for cross-mapping errors as noted above.

We investigated one GTEx trans-eQTL in greater detail for illustration – variant: chr5:149826526 and gene: RP11-343H5.4 (ENSG00000224114) – which was significant in each of 5 GTEx tissues. RP11-343H5.4 is a pseudogene on chromosome 1. In the coverage plots of the gene, we noticed that reads were aligned to only a fraction of the exonic region of the gene; if the gene were truly expressed, we would expect reads being mapped across the whole exon (Figure 2.7). RP11-343H5.4 is cross-mappable with RPS14 (ENSG00000164587), a protein-coding gene in chromosome 5 near the putative trans-eQTL variant. There was also a cis-association between the variant and RPS14. *k*-mers from RPS14 indeed map to the region within RP11-343H5.4, where we observed a non-zero number of reads. Interestingly, in this case, read mapping appears to be genotype-dependent - the variant at chr5:149826526 alters sequence such that it would lead to reads from RPS14 uniquely, but likely incorrectly, mapping to RP11-343H5.4.

Finally, we found that cross-mappable eQTLs, which we believe to be

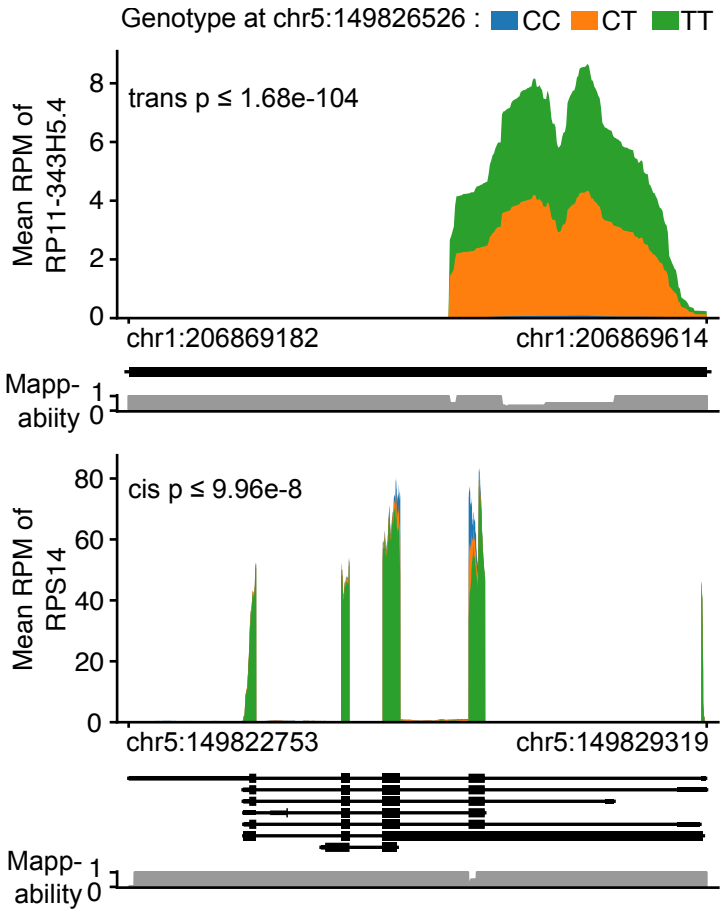

Figure 2.7: An example of likely false positive trans association between the variant chr5:149826526 and the gene RP11- 343H5.4. The coverages (reads per million, RPM) of the trans-eGene RP11-343H5.4 (top) and its cross-mapping gene RPS14 (bottom) in Thyroid are shown along with their exons and UTRs (black lines below the coverage plot), and mappability of 75-mers. The regions of mappability less than 1.0 have sequence similar between the two genes.

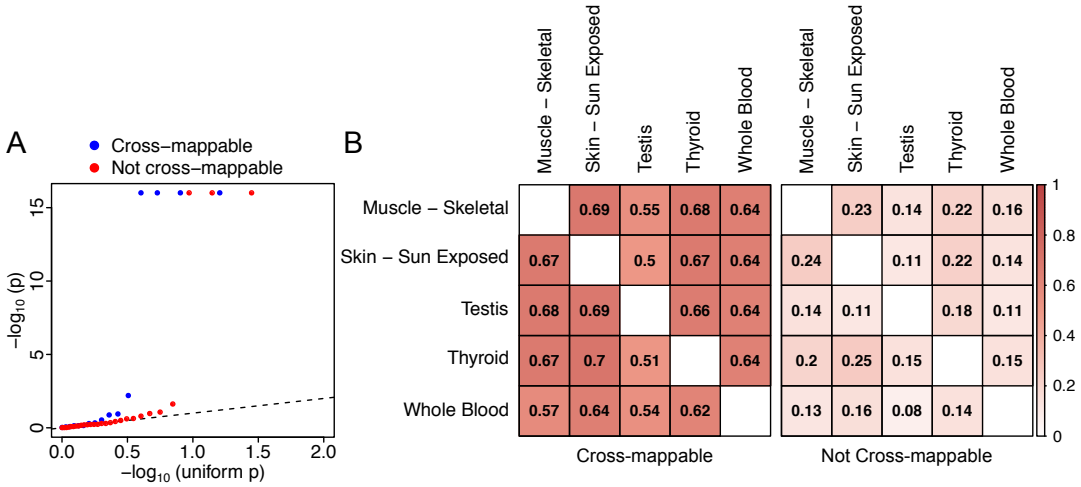

Figure 2.8: Trans-eQTL replication. (A) Q-Q plot, replication p-values from DGN for variant-gene pairs discovered in GTEx Whole Blood, grouped by cross-mappability. (B) The fraction of significant eQTLs in each GTEx tissue (row) replicated in another tissue (column) at FDR ≤ 0.05, for cross-mappable eQTLs (left) and not cross-mappable eQTLs (right).

enriched for false-positives, are highly replicable between datasets. This misleading replication occurs because it is driven by the underlying sequence of the genome, and similar alignment errors frequently occur regardless of tissue and study. We showed this by measuring the replication between the significant trans-eQTLs detected at FDR ≤ 0.05 from whole blood from GTEx and whole blood data from the DGN study (Battle et al., 2014). To avoid the effects of linkage disequilibrium, we tested for trans-association in DGN only for the best variant per GTEx trans-eQTL gene (with the lowest p-value in GTEx), where both the variant and the gene were present in the DGN data. At FDR ≤ 0.05, only 10.71% (3 out of 28) non-cross-mappable trans-eQTLs were replicated in DGN while 31.25% (5 out of 16) cross-mappable trans-eQTLs

were replicated. The Q-Q plot in Figure 2.8A shows that cross-mappable trans-eQTLs were more likely to be replicated compared to non-cross-mappable ones. We observed the same phenomenon when we attempted to replicate significant trans-eQTLs detected from one GTEx tissue in other GTEx tissues. On average, 63.0% (range: 50.3-70.2%) and 16.3% (range: 7.6-25.1%) of cross-mappable and non-cross-mappable trans-eQTLs, respectively, were replicated (Figure 2.8B). This suggests that replication of a trans-eQTL does not necessarily indicate a true positive. Overall, we suggest that regardless of replication, cross-mappable trans-eQTLs require further investigation to establish that they arise from biological regulation rather than alignment artifacts.

2.3.2 Effect of cross-mappability in co-expression analysis

Next, we evaluated evidence that alignment errors between genes can cause spurious correlation between gene expression levels (co-expression). If alignment errors did not affect co-expression analysis, we would expect that the distribution of pairwise correlation between cross-mappable genes would not deviate from that between non-cross-mappable genes. To test this, we used the gene expression data in five tissues from GTEx v7 after correction for covariates and latent confounders (see Methods). For each tissue, we selected a random set of 10,000 non-cross-mappable gene pairs and a random set of 10,000 cross-mappable gene pairs chosen with probability proportional to their cross-mappabilities (sampling probability proportional to cross-mappability ensures sampling from the whole cross-mappability range, as opposed to just from the massive number of low cross-mappability pairs). Then we computed

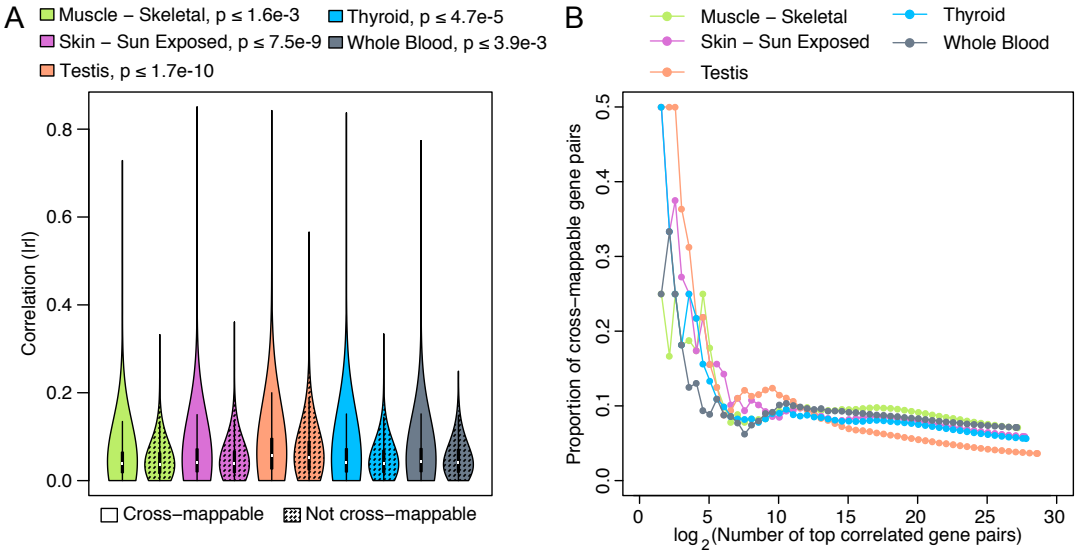

Figure 2.9: Effect of cross-mappability on co-expression. (A) Comparison of co-expression between randomly drawn pairs of cross-mappable genes and not cross-mappable genes. Each violin plot shows the distribution of the absolute Pearson correlation (y-axis) between corrected gene expression levels of randomly drawn 10,000 gene pairs in a tissue (color). P-value of the Wilcoxon test to determine whether cross-mappable genes are more correlated than not cross-mappable genes in each tissue is shown in the legend. (B) Fraction of top co-expressed genes that are cross-mappable and thus potential false positives.

the absolute Pearson correlation ($|r|$) between expression levels of the genes in each randomly selected pair. We found that expression levels of cross-mappable genes were more correlated than expression levels of non-cross-mappable genes (median p across tissues $\leq 4.7 \times 10^{-5}$, Wilcoxon rank-sum test, Figure 2.9A). The difference was more significant when uncorrected data were used (median $p \leq 1.3 \times 10^{-74}$) We also observed that the correlation coefficient tends to increase with increasing levels of cross-mappability between genes (Figure 2.10), indicating a high rate of false co-expression in the most highly cross-mappable genes. The increased correlation between cross-mappable genes was observed even after discounting genes from same gene family (Figure 2.11), somewhat alleviating concerns that our observations were due to exclusively true functional relationships. We observed a similar pattern using data from an independent RNA-seq study, DGN (Figure 2.12).

To demonstrate the impact of this pattern on a realistic genome-wide co-expression analysis, we evaluated how many of the top-most correlated gene pairs in each GTEx tissue suffer from cross-mappability. We observed that cross-mappable pairs of genes are over-represented among the top hits, with gene pairs ordered by the absolute Pearson correlation after excluding pairs of genes whose genomic coordinates actually overlap (Figure 2.9B). Overall, the impact of cross-mappability on co-expression appears to be less than on trans-eQTL analysis, but the phenomenon may still require consideration when examining specific co-expressed gene pairs or enrichment patterns.

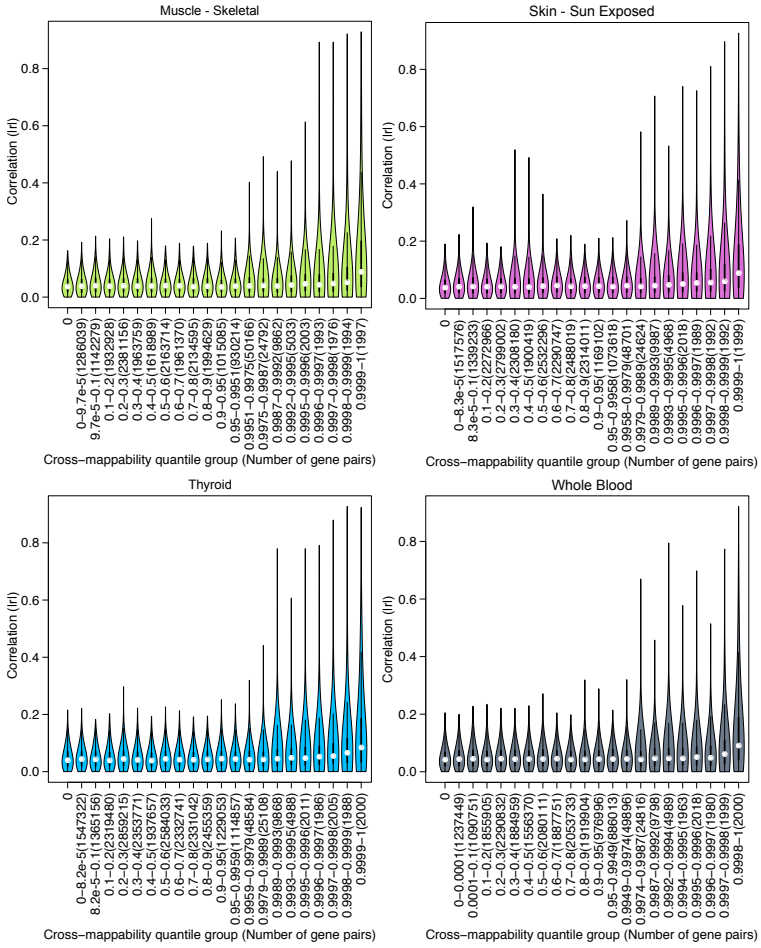

Figure 2.10: Correlation between random gene pairs increases with cross-mappability. Gene pairs available in each tissue were categorized into 22 groups (x-axis) based on quantiles. A quantile group "$q_1 - q_2(n)$" represents gene pairs of $(q_1 * 100, q_2 * 100]$-th percentile of cross-mappability with a total of n pairs. In order to visualize the impact of the highest range of cross-mappability, the rightmost nine quantile groups were selected in such a way that each contains about a certain number of pairs: (from right) 2,000, 2,000, 2,000, 2,000, 2,000, 5,000, 10,000, 25,000, 50,000. The leftmost quantile group "0" represents gene pairs which are not cross-mappable. From each group, 1,000 gene pairs were randomly selected where the probability of drawing a pair was proportional to its cross-mappability. Each violin plot shows the distribution of absolute Pearson correlation (y-axis) between corrected expressions of the genes in each pair.

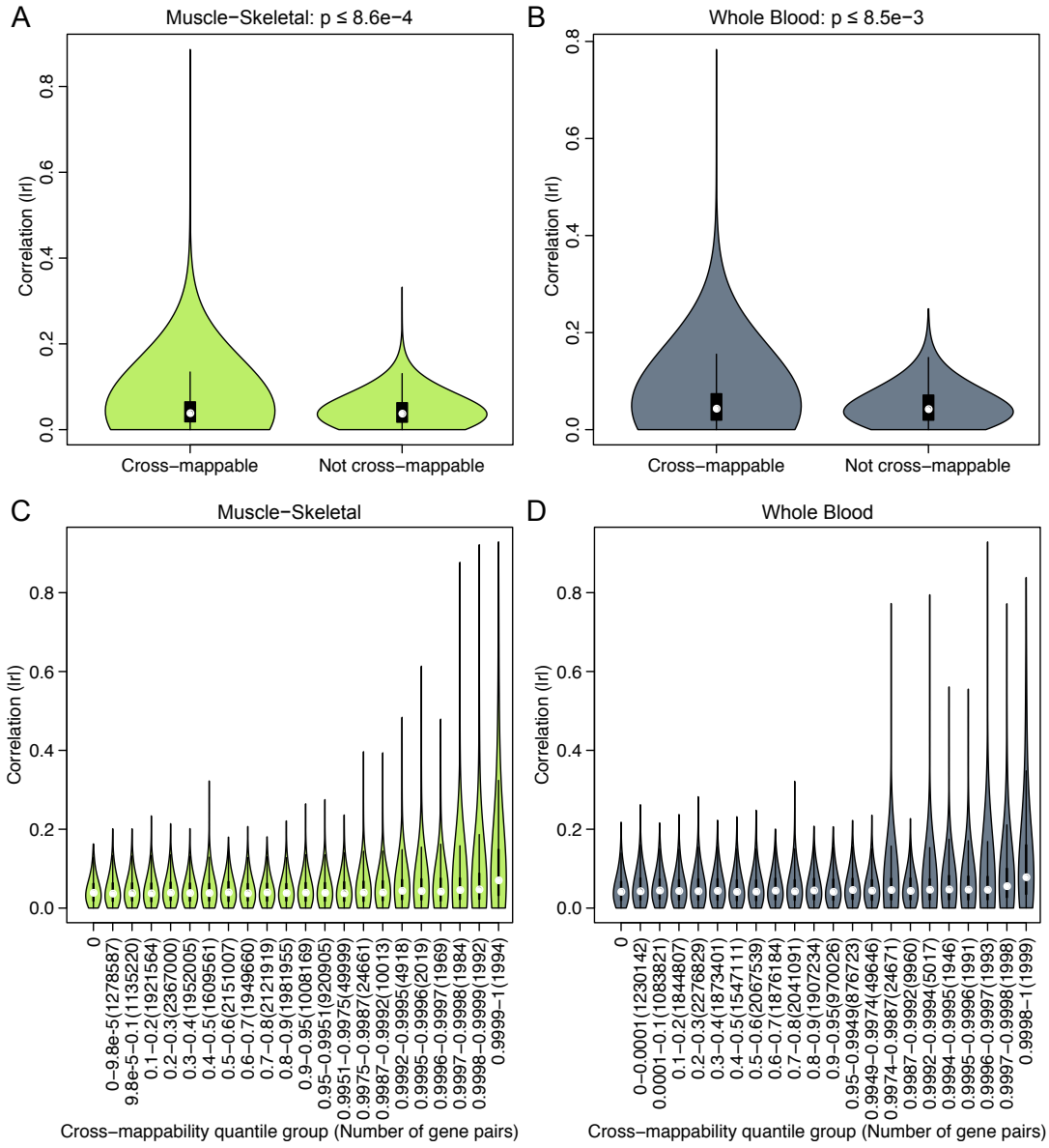

Figure 2.11: Increased correlation between cross-mappable genes is not exclusively due to sequence similarity between genes from same gene family. Here, two genes in the same HGNC gene family were artificially excluded from cross-mappable pairs. We computed the absolute Pearson correlation between gene pairs within different groups as described in Figure 2.9A and Figure 2.10. Note: gene family information was downloaded from www.genenames.org. A-B) Comparison of co-expression between 10,000 randomly drawn pairs of cross-mappable and not cross-mappable genes in Muscle – Skeletal (A) and Whole Blood (B). C-D) Random correlation between genes in Muscle – Skeletal (C) and Whole Blood (D).

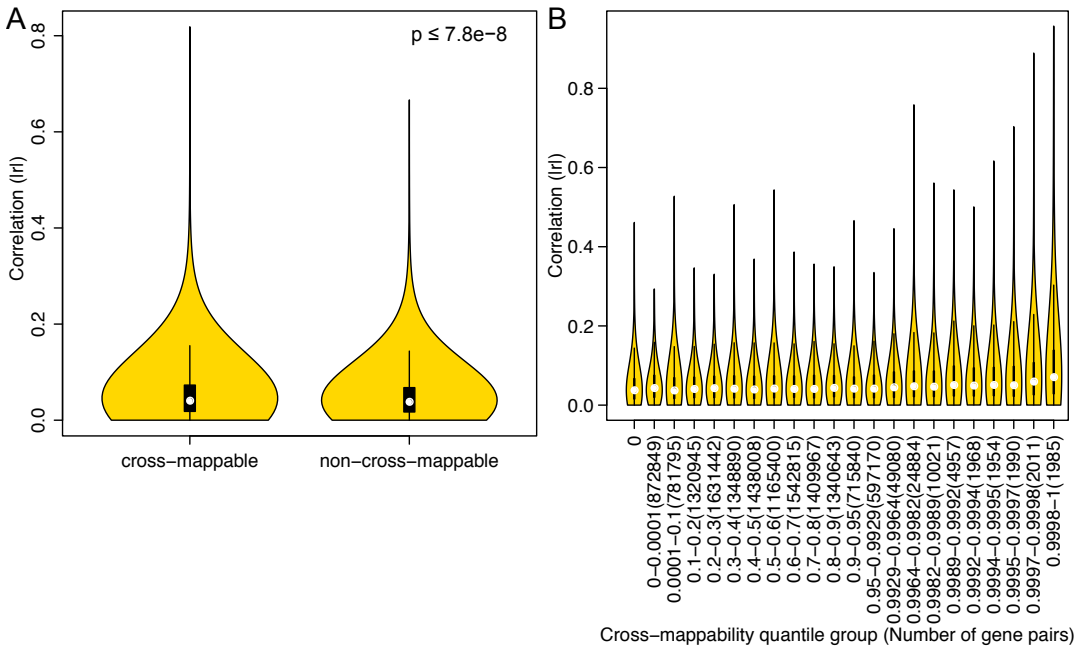

Figure 2.12: Co-expression analysis using gene expression data from DGN. A) Comparison of co-expression between 10,000 randomly drawn cross-mappable and non-cross-mappable gene pairs. B) Random correlation between genes in DGN increases with cross-mappability.

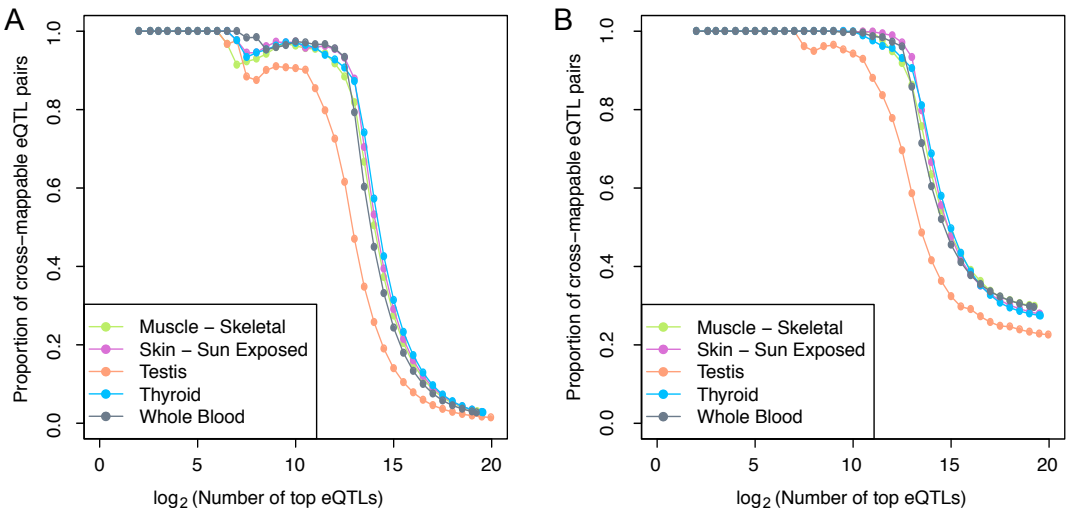

Figure 2.13: Effects of varying k-mer length and the number of mismatches allowed. Cross-mappability among the top GTEx trans-eQTLs when A) 75-mers (instead of 36-mers) from UTRs were used, B) a maximum of 3 (instead of 2) mismatches were allowed. 67.2% and 76.1% of the significant trans-eQTLs remain cross-mappable in (A) and (B), respectively, compared to 75.14% using 75-mers from exons and 36-mers from UTRs with 2 mismatches in the original analysis. In both cases, cross-mappable trans-eQTLs still tend to be the most highly significant.

2.3.3 Impact of alternative quantification and parameter settings

We have made several versions of our cross-mappability resources publicly available for the human genome (hg19 and GRCh38) (Saha & Battle, 2019b), and also published code in Github (Saha & Battle, 2019a). Researchers should carefully choose settings according to the study design and goals. Genome version and gene annotations can be directly matched, but other parameter choices such as k and the maximum number of mismatches allowed in alignment may affect the detection of false positives. Small values of k will produce more conservative cross-mappability scores, but large k may not correctly handle small exons or UTRs. For example, if 75-mers (instead of 36-mers) were used from UTRs, a smaller proportion of trans-eQTLs (67.2% instead of 75.14%) would appear as cross-mappable in GTEx, although cross-mappable trans-eQTLs would still tend to be most highly significant (Figure 2.13A). Similarly, increasing the number of mismatches allowed in k-mer alignment results in an increased number of cross-mappable trans-eQTLs (Figure 2.13B). For convenience, k and the number of mismatches are configurable in our software so that, if needed, one can compute cross-mappability scores with settings appropriate for a given study.

We also note that utilization of improved alignment and quantification methods to generate gene expression data may also be helpful to avoid false positives. For example, quantification of gene expression levels using RSEM (Li & Dewey, 2011), an expectation maximization based quantification tool, results in a smaller fraction of false positive trans-eQTLs (60.17%) than that

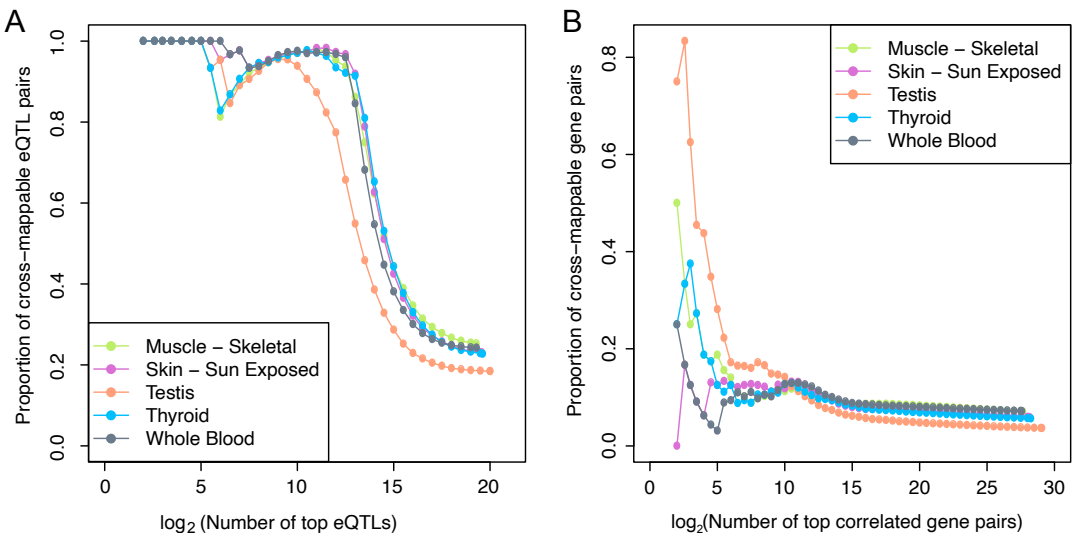

Figure 2.14: Effects of EM-based quantification methods. A) We computed trans-eQTLs using RSEM-quantified data. A total of 27,035 trans-eQTLs were detected at FDR $\leq$ 0.05, 60.17% of which were cross-mappable compared to 75.14% with RNA-SeQC-quantified data. The plot shows the fraction of cross-mappable trans-eQTLs among the top significant variant-gene pairs (ordered by increasing FDR) in each tissue (color). Here, we observed a modest improvement by RSEM. B) Fraction of top co-expressed genes that are cross-mappable and thus potential false positives. Cross-mappable gene pairs still appear abundant in most correlated genes of multiple tissues.

using RNA-SeQC (75.14%). However, potential false positives due to cross-mappability still remain abundant in both trans-eQTL and co-expression studies (Figure 2.14).

2.4 Discussion

Misalignment of short sequencing reads has the potential to induce false positives in association studies. For RNA-seq, both trans-eQTL and co-expression analyses are susceptible to these artifacts, related to false positives in microarray analysis due to probe cross-hybridization. This is readily apparent from the enrichment of processed pseudogenes among the top hits for such association studies, but misalignment can affect protein-coding genes as well. Our results demonstrate that trans-eQTL associations in a standard pipeline are dominated by potential false-positives due to sequence similarity and replication rates between studies may be artificially inflated due to this pattern. Additionally, genes with sequence similarity display more correlated expression levels, and mapping errors should be considered in co-expression analysis as well.

Our results do not imply that all instances of co-expression or trans-eQTL associations arising from genes with sequence similarity are in fact false positives. Genes with sequence similarity also sometimes have true functional relationships. Pseudogene transcripts may interact with coding transcripts, and some associations with pseudogene expression may reflect true regulatory relationships (Pink et al., 2011). Furthermore, the background (random) rate of sequence similarity between any two regions in the human genome

is above zero; that is, a hit may occur between regions of sequence similarity by chance, even when no actual misalignment of reads has taken place. However, we believe the exceedingly high fraction of cross-mappable regions among trans-eQTLs from a naive analysis warrants suspicion that these hits are predominantly false positives. Researchers should consider their particular application and tolerance for false negatives and false positives when applying filters targeting alignment errors. Other information, such as base-level coverage plots and outside functional information can help disambiguate particular cases of interest.

Extensions, modifications, and other approaches related to this problem should also be considered. First, specifics of study design, and in particular sequencing read length, should be taken into account when using our data to filter potential false positives. If read length is much shorter or longer than our *k*-mer setting, our existing data may be insufficient and new mappability and cross-mappability estimates should be derived. In the initial resource provided, we used *k*-mer alignment to the genome, which does not directly handle splice junctions in transcriptomic data (and also limits appropriate *k*-mer length even for studies with longer reads). Alignment to the transcriptome or splice-aware alignment may offer future improvements, but computational cost and inaccuracies due to incorrect annotation will have to be evaluated. Our observations and methods may be relevant to analyses of other functional genomic data as well, including detection of interactions from HI-C, and detection of associations with data types such as ATAC-seq or ChIP-seq. Other approaches, such as filtering reads themselves before

quantification can also be applied if raw reads rather than quantified data are available and tractable (van de Geijn et al., 2015).

Our evaluation provides evidence that misalignment of reads should be considered as a potential source of false positives in association studies, particularly for trans-eQTL analysis. The resources we provide can be used directly to filter potential false positives, or the ideas presented may be tuned and adapted to new studies and data types.

2.5 Data and code availability

Pre-computed cross-mappability resources for human genomes (hg19 and GRCh38) are available on FigShare.

Github repository to compute cross-mappability: https://github.com/battle-lab/crossmap.

Github repository to replicate analyses in the manuscript: https://github.com/battle-lab/crossmap_analysis.

Chapter 3

Joint regulation of transcription and alternative splicing

In the previous chapter, we reported a source of false positives in transcription regulation studies. In this chapter, we present our work on joint regulation of *transcription* and *alternative splicing*.

Transcription is the process by which a double-stranded DNA is copied (*transcribed*) to produce single-stranded precursor messenger RNAs (pre-mRNAs) containing codes for proteins (Figure 1.3). It determines the amount of RNA and consequently the amount of protein produced from each gene in a given context. The total amount of RNA generated from a gene is generally called *total expression* (TE) or simply *expression* of the gene. Every organism needs to maintain proper gene expression levels to perform necessary activities.

RNA splicing is an important process in which introns are removed from each pre-mRNA and exons are joined together to form a *mature mRNA* or simply *mRNA*. By varying the composition of exons, the splicing process

may produce different mature mRNAs from the same pre-mRNA. This phenomenon is known as *alternative splicing* (Figure 1.3). Each type of mature mRNA produced from the same gene through alternative splicing is called an *isoform*. Importantly, each isoform contains a different sequence and generally produces different types of proteins. In fact, alternative splicing is an essential mechanism in complex organisms, including humans, to produce many different proteins for all necessary tasks. A mis-splicing can effectively alter downstream proteins causing diseases. According to one study, about one-third of all disease-causing mutations alter RNA splicing (Lim et al., 2011). According to another study (Kahles et al., 2018), alternative splicing events are up to 30% more common in tumor samples compared to control samples, underscoring the importance of understanding the regulation of alternative splicing.

While a few splicing factors are known, specific regulatory genes involved in splicing regulation remain poorly understood relative to transcription (Melé et al., 2015; Scotti & Swanson, 2015). In this project, led by me, we extended the framework of co-expression networks to jointly study transcription and splicing and the interplay between them. Our main contributions are the following:

- We developed a framework called *Transcriptome-Wide Networks (TWNs)* for combining total expression and relative isoform levels into a single sparse network, capturing the interplay between the regulation of splicing and transcription.

- We built TWNs for sixteen human tissues, and found that hubs in these

networks were strongly enriched for splicing and RNA binding genes, demonstrating their utility in unraveling regulation of splicing in the human transcriptome.

- Next, in collaboration with the Engelhardt Lab at Princeton University, we used a Bayesian biclustering model that identifies network edges unique to a single tissue to reconstruct Tissue-Specific Networks (TSNs) for 26 distinct tissues.

- Finally, we found genetic variants associated with pairs of adjacent nodes in our networks, supporting the estimated network structures and identifying 20 genetic variants with distant regulatory impact on transcription and splicing.

This work was published in Genome Research (Saha et al., 2017), and this chapter is based on the published article.

3.1 Introduction

Gene co-expression networks are an essential framework for elucidating gene function and interactions, identifying sets of genes that respond in a coordinated way to environmental and disease conditions, and highlighting regulatory relationships (Penrod et al., 2011; Xiao et al., 2014; Yang et al., 2014). Each edge in a co-expression network reflects a correlation between two transcriptional products, represented as nodes (Stuart et al., 2003). Most gene co-expression networks focus on correlation between total gene expression

levels, with edges representing transcriptional co-regulation. However, post-transcriptional modifications, including alternative splicing, are important in creating a transcriptome with diverse biological functions (Matlin et al., 2005). Mutations that lead to disruption of splicing play an important role in tissue- and disease-specific pathways (DeBoever et al., 2015; Lee et al., 2012; Li et al., 2016d; López-Bigas et al., 2005; Wang et al., 2008; Ward & Cooper, 2010).

While a number of splicing factors are known, regulation of splicing and specific regulatory genes involved remain poorly understood relative to the regulation of transcription (Melé et al., 2015; Scotti & Swanson, 2015). Although abundance of different isoforms can be influenced by processes including usage of alternative transcription start or end sites and RNA degradation, variation in isoform levels is often the direct result of alternative splicing. RNA sequencing now allows quantification of isoform-level expression, providing an opportunity to study regulation of splicing using a network analysis. However, current research estimating RNA isoform-level networks (Li et al., 2016a; Li et al., 2015; Li et al., 2014) has focused on total expression of each isoform, and the resulting network structures do not distinguish between regulation of transcription and regulation of splicing in an interpretable way. Initial work on clustering relative isoform abundances has also been explored (Dai et al., 2012; Iancu et al., 2015), but does not support discovery of fine-grained network structure or identification of regulatory genes. Neither approach has been applied to large RNA-seq studies for network reconstruction in diverse tissues.

Another important gap in our interpretation of regulatory effects in complex traits is a global characterization of co-expression relationships that are only present in a specific tissue type. Per-tissue networks have been estimated for multiple tissues (Pierson et al., 2015; Piro et al., 2011), but, critically, these analyses do not directly separate effects unique to each tissue from shared effects found in all or many tissues. Recent studies have recognized the essential role that tissue-specific pathways play in disease etiology (Greene et al., 2015), but have developed these per-tissue networks by aggregating single tissue samples across multiple studies. However, differences in study design, technical effects, and tissue-specific expression make cross-study results difficult to interpret mechanistically, with large groups of genes expressed in similar tissues and studies tending to be highly connected rather than including sparse edges that detail tissue-specific network structure (Lee et al., 2004).

In this work, we reconstruct co-expression networks from the Genotype Tissue Expression (GTEx) v6 RNA sequencing data (The GTEx Consortium, 2015, 2017), including 449 human donors with genotype information and 7,310 RNA sequencing samples across 50 tissues. We apply computational methods designed to reveal novel relationships between genes and across tissues as compared to previous analyses, specifically addressing two important goals in regulatory biology: identification of edges reflecting regulation of splicing, and discovery of edges arising from gene relationships unique to specific tissues. We introduce a new framework, *Transcriptome Wide Networks (TWNs)*, which capture gene relationships that reflect regulation of alternative splicing in an interpretable model. We built TWNs to identify candidate

regulators of both splicing and transcription across sixteen tissues. Next, we identified *Tissue-Specific Networks (TSNs)* for 26 tissues, where each network edge corresponds to a correlation between genes that is uniquely found in a single tissue. We study the biological interpretation of both network types by quantifying enrichment of known biological functions among well-connected nodes. Finally, we use genetic variation to validate network edges from each network by testing associations between a regulatory variant local to one gene with that gene's network neighbors. Interpretation of regulatory and disease studies will benefit greatly from these networks, providing a much more comprehensive description of regulatory processes including alternative splicing across diverse tissues.

3.2 Methods

3.2.1 Transcriptome-wide network (TWN)

We developed a method to estimate Transcriptome-Wide Networks (TWNs) from RNA-seq data that captures the co-regulation of alternative splicing, in addition to co-expression, across multiple genes.

We first quantified both total expression level (TE) and isoform expression levels of each gene in each RNA-seq sample and then computed isoform ratios (IR), representing the relative abundance of each isoform with respect to the total expression of the gene (Figure 3.1A). Unlike a traditional co-expression network that includes only one type of nodes generally corresponding to total expression levels, we included two types of nodes in a TWN, corresponding to total expression levels and isoform ratio nodes of each gene. By including

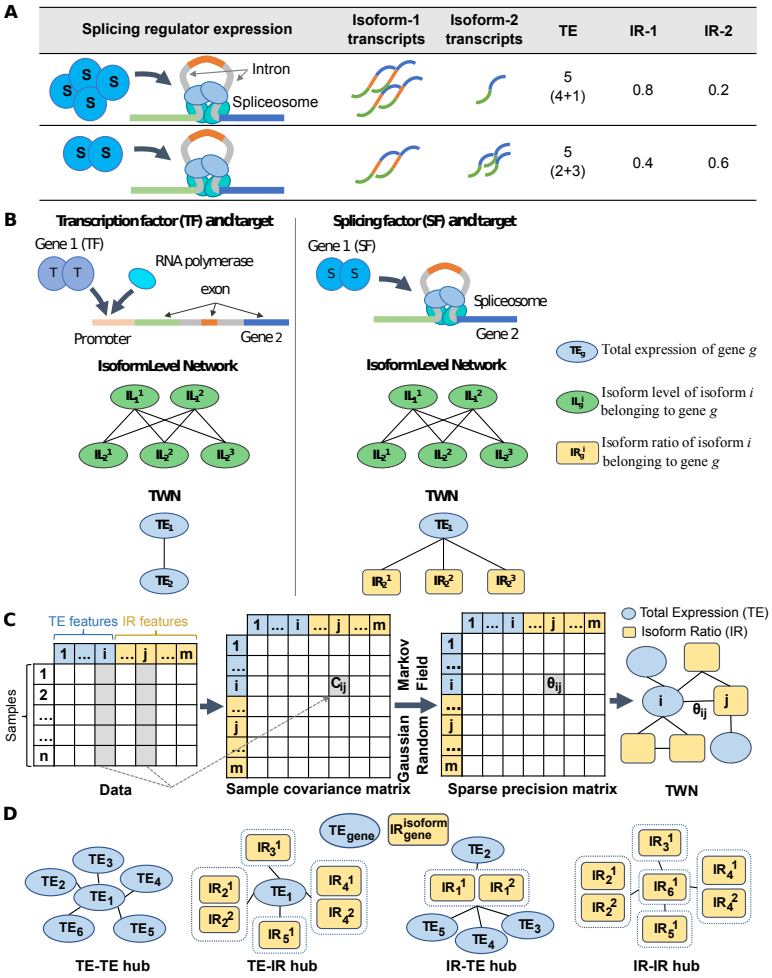

Figure 3.1: Transcriptome-Wide Network conceptual framework. (A) Schematic of the effect of a splicing regulator on inclusion of a cassette exon, and resulting total expression and isoform ratios of the target gene. Higher expression of a splicing regulator S (first row) results in relatively more transcripts of isoform-1 and fewer of isoform-2. Total expression level is constant (5), but isoform ratios are different (0.4 and 0.6) as splicing factor S levels change. (B) Left: Standard isoform-level network and TWN representation when a transcription factor (Gene 1 with 2 isoforms) regulates transcription of its target (Gene 2 with 3 isoforms). Right: Standard isoform-level network and TWN representation when a splicing factor (Gene 1 with 2 isoforms) regulates splicing of its target (Gene 2 with 3 isoforms). (C) We created a sparse precision matrix Θ (inverse covariance) from the sample covariance matrix using graphical lasso. A non-zero value (Θ_{ij}) in the precision matrix denotes an edge between i-th feature and j-th feature in the network. (D) Different types of hubs in a TWN. Dotted rectangles group together isoform ratios for different isoforms of the same gene.

information about isoforms, a TWN can capture splicing regulation, which a traditional co-expression network cannot.

A TWN is also critically different from a standard correlation network of isoform expression levels (not isoform ratios) to distinguish correlation due to transcription and alternative splicing. An isoform-level network cannot distinguish an edge due to transcription from an edge due to splicing, because both transcription and splicing affect the expression level of an isoform. In contrast, splicing affects isoform ratio, but not total expression, and transcription affects the total expression level, but not isoform ratio. A TWN offers an interpretable framework to distinguish between transcription and splicing, and to reveal the interplay between both mechanisms. For example, to represent the relationship between a transcription factor (TF) and expression of a target gene, where all isoforms are equally affected, a standard isoform-level network would require edges from each isoform level of the TF to each isoform level of the target (Figure 3.1B). The same structure would also represent the relationship between a splicing factor (SF) and its target gene where the transcription is grossly unaffected but the relative production of isoforms is altered.

Taking both total expression levels and isoform ratios as features, we estimated a sparse precision matrix (Θ) from the sample covariance matrix (S) using a graphical lasso (Friedman et al., 2008) approach (Figure 3.1C). A non-zero entry in the sparse matrix implies that the corresponding two nodes are correlated after controlling for other nodes. We modified the standard graphical lasso to penalize different types of edges by different weights. Specifically,

we optimized the following objective:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} - \log\det\ \Theta + tr(S\Theta) + \|\Lambda \circ \Theta\|_1 \tag{3.1}$$

where the entry in r-th row and c-th column of the penalty matrix Λ was

$$\Lambda_{rc} = \begin{cases} \lambda_d & \text{if } r = c \\ \lambda_s & \text{if } r \neq c \text{ and } gene(r) = gene(c) \\ \lambda_{tt} & \text{if } gene(r) \neq gene(c) \text{ and } type(r) = type(c) = \text{`TE'} \\ \lambda_{ti} & \text{if } gene(r) \neq gene(c) \text{ and } \{type(r), type(c)\} = \{\text{`TE',`IR'}\} \\ \lambda_{ii} & \text{if } gene(r) \neq gene(c) \text{ and } type(r) = type(c) = \text{`IR'}. \end{cases} \tag{3.2}$$

Here, $type(k)$ denotes whether or not the k-th feature represents total expression ('TE') or isoform ratio ('IR'). $gene(k)$ denotes the gene that the k-th feature belongs to.

We did not penalize diagonal entries ($\lambda_d = 0$), and we put a small non-zero penalty ($\lambda_s = 0.05$) for edges between distinct features belonging to the same gene, such as distinct isoforms of the same gene. We selected the other penalties ($\lambda_{tt}, \lambda_{ti}, \lambda_{ii}$) in such a way that the network had a scale-free topology with a reasonable number of edges.

3.2.2 Tissue-specific network (TSN)

A per-tissue TWN contains both shared and tissue-specific co-expression relationships between genes, without making any distinction between them, reflecting the full gene network in each tissue. To directly assess the tissue-specificity of co-expression relationships, we applied a Bayesian biclustering

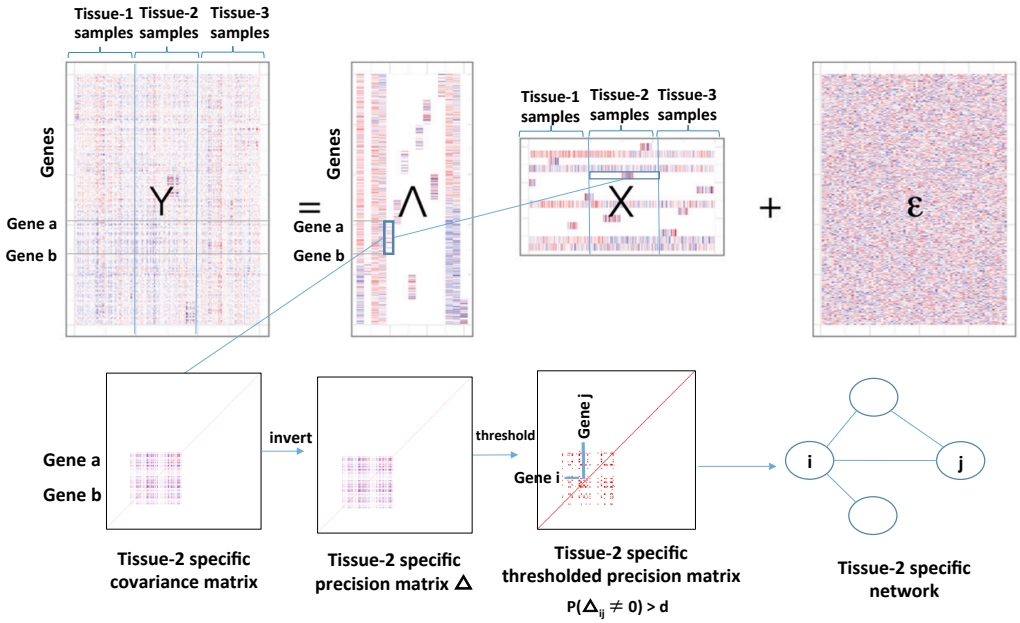

Figure 3.2: Tissue-specific network (TSN) conceptual framework. BicMix, a Bayesian sparse factor analysis based model, decomposes a gene-by-sample expression matrix (Y) into a gene-by-K loading matrix (Λ), a K-by-sample factor matrix (X), and a gene-by-sample residual matrix (ϵ), where K is the number of latent factors. BicMix induces sparsity in both Λ and X, and thus identifies clusters of co-expressed genes that are co-expressed in a subset of samples. Using the gene loadings corresponding only to factors with non-zero values in a single tissue, a precision submatrix (Δ) corresponding to the non-zero genes can be estimated; standardized, these values correspond to partial correlation. Thresholding these partial correlations using FDR, each non-zero value corresponds to an edge between a pair of genes in the tissue-specific gene co-expression network. By estimating the gene covariance matrix using only components with non-zero values among the tissue of interest in BicMix, we explicitly remove all covariation that is found outside of the tissue of interest. This shared covariation may also include covariation due to batch effects, population effects, cross-tissue expression QTLs, or cellular housekeeping pathways; while this shared variation is captured in the BicMix model, it is ignored when building the TSNs.

framework, BicMix (Gao et al., 2016), and reconstructed tissue-specific networks (Figure 3.2). BicMix uses a sparsity-inducing prior to differentiate between gene co-expression relationships specific to a single tissue and those shared across tissues, simultaneously controlling for batch effects, population structure, and shared individual effects across tissues (Gao et al., 2016).

3.2.3 Data from GTEx project

RNA-seq data: We collected RNA sequencing and genotyping data from the Genotype-Tissue Expression (GTEx) consortium (The GTEx Consortium, 2015). GTEx obtained tissue samples (averaging about 28 per individual) from postmortem donors, between ages 21 and 70, BMI 18.5 to 35, and not under exclusionary medical criteria such as whole-blood transfusion within 24 hours or infection with HIV. 76 base pair (bp) pair-ended mRNA sequencing was performed with Illumina HiSeq 2000 following the TrueSeq RNA protocol, resulting approximately 50 million reads per sample. After quality controlling, we aligned the RNA-seq reads using the STAR aligner (Dobin et al., 2013) in 2-pass mode with GENCODE v.19 annotation retaining only uniquely mapping reads. We then performed transcript and gene quantification using RSEM v1.2.20 (Li & Dewey, 2011). We used RNA-seq data across 50 tissues in 449 individuals.

Genotype data: Approximately 1.9 million SNPs were genotyped using whole blood samples with Illumina HumanOmni 2.5M and 5M BeadChips. Additional variants were imputed using IMPUTE2 (Howie et al., 2009). The genotypes were filtered for MAF $\geq$ 0.05, leaving approximately 6 million

variants.

3.2.4 Pre-processing for per-tissue TWNs

We considered only protein-coding genes on the autosomes and Chromosome X to construct TWNs in all tissues. We used genes and isoforms with at least 10 samples with ≥ 1 TPM and ≥ 6 reads. We filtered out genes where the Ensembl gene ID did not uniquely map to a single HGNC gene symbol. Isoform ratio was computed by using annotated isoforms in GENCODE V19 annotation, and undefined isoform ratios ($0/0$, when none of the isoforms was expressed) were imputed from the mean ratio per isoform across individuals. Each gene's least abundant isoform was excluded to avoid linear dependency between isoform ratio values. We log-transformed the total expression data and standardized both total expression levels and isoform ratios. To correct hidden confounding factors, we applied HCP (Hidden covariates with prior) (Mostafavi et al., 2013), whose parameters were selected based on an external signal relevant to regulatory relationships. Namely, we selected parameters that produced maximal replication of an independent set of trans-eQTLs from meta-analysis of a large collection of independent whole blood studies (Westra et al., 2013). For both total expression levels and isoform ratios of genes in all tissues, the best HCP parameters ($k = 10$, $\lambda = 1$, $\sigma_1 = 5$, $\sigma_2 = 1$), which consistently reproduced a largest subset of the gold-standard trans-eQTLs in GTEx *whole blood* samples even when subsetting the number of samples, were used for correcting data. Finally, quantile-normalization to a standard normal distribution was applied.

To avoid spurious associations due to mis-mapped reads, we filtered out genes with mappability (Saha & Battle, 2018) < 0.97 and their isoforms. We also filtered out isoforms of a gene if the mean IR of the most dominant isoform was ≥ 0.95.

For computational tractability, we selected 6,000 genes and 9,000 isoforms in each tissue from available genes and isoforms that passed other filtering steps. To do so, we first considered genes or isoforms if > 10 samples have TPM > 2 or reads > 6. To obtain the final set of genes, we first considered the top 9,000 genes based on their average expression levels and then selected the top $6,000$ highly variable genes across individuals. Similarly, to obtain the final set of isoforms, we first considered the $13,500$ genes with the highest expressed isoform levels on average. We reduced this to 11,250 genes based on the entropy of isoform ratios across individuals, normalized by the maximum entropy possible with the same number of isoforms, and finally took the top 9,000 most highly variable isoforms in terms of TPM values. On average, the finally selected isoforms for each tissue belong to 4,357 unique genes.

3.2.5 Pre-processing for TSNs

We normalized the gene level TPM data for GC content, length, and depth. For each tissue, we removed genes that had zero read counts in more than 90% of samples. We took the intersection of all remaining genes across the 50 tissues, and only used those 15,589 genes for the analysis. All 50 tissue expression matrices were appended together and subsequently quantile normalized within each gene across all tissues.

3.2.6 TWN hub ranking

We ordered the network hubs by degree centrality for each tissue according to the number of unique gene-level connections to avoid the effect of different number of isoforms per gene. To do this, we created a gene-level network from the original TWNs by keeping TE nodes as they were and grouping all isoforms of the same gene together to form a compound IR node. We put an edge between a compound IR node and a TE node (or another compound TE node) if any isoform of the compound had an edge with the TE node (or any isoform of the other compound) in the original TWN, and the weight was equal to the sum of absolute weights of all such edges in the original TWN. TE-TE and IR-TE hubs were ordered by the number of TE nodes they were connected with. TE-IR and IR-IR hubs were ordered by the number of compound IR nodes they were connected with. If multiple hubs had the same number of connections, ties were broken by the sum of corresponding edge weights.

3.2.7 TF-target enrichment in TE-TE edges of TWN

We downloaded transcription factors (TFs) and their known targets from ChEA (Lachmann et al., 2010). We measured the number of known TF-target relationships captured by a network, i.e., a TF and its target's total expression nodes were directly connected with each other. We generated the null distribution of the number of known TF-target relationships by computing the same test statistics for random networks, generated by permuting gene names among network nodes 1000 times. Then, we computed the empirical p-value

as the proportion of those iterations for which the random network had at least as many known TF-Target edges as the test network. We fitted a Weibull distribution on the log(1+fraction of known TF-Target edges) to quantify the p-values.

3.2.8 TWN hubs specific to a group of related tissues

To find hubs specific to a group of tissues, we used rank-product to rank hubs in both the target group of tissues, and all other tissues, separately. Then, we normalized ranks so that the top- and bottom-ranked hubs have a score of 1 and 0, respectively. Let the normalized rank of a gene in the target group of tissues and other tissues be r_t and r_o, respectively. Then, the F-score for the gene (r),

$$r = \frac{2}{\frac{1}{r_t} + \frac{1}{1-r_o}}, \tag{3.3}$$

will be high if it ranks high in the target group, but low in other tissues.

We computed related tissue specific hubs for five groups of related tissues: 1) *skin– sun exposed* and *skin – not sun exposed*, 2) *adipose – subcutaneous*, *adipose – visceral* and *breast – mammary*, 3) *heart – left ventricle* and *skeletal muscle*, 4) *esophagus – mucosa* and *esophagus – muscularis*, and 5) *artery – aorta* and *artery – tibial*.

Cis-eQTLs from TSNs

For each tissue in which we recovered a TSN, we used the same set of genes and expression values as described for TSN creation, prior to taking the

intersection of genes across all tissues. PEER factors were used to quantify effects of unobserved confounding variables (Stegle et al., 2012). We optimized the number of PEER factors by tissue to a test chromosome (Chromosome 11) to maximize the number of identified cis-eQTLs. The linear model of Matrix-eQTL (Shabalin, 2012) was used to test all SNPs within the 100 kb window of a gene's transcription start site (TSS) or end site (TES) using an additive linear model. We included in association mapping a tissue-specific number of PEER factors, sex, genotyping batch, and three genotype principal components. The correlation between SNP and gene expression levels was evaluated using the estimated t-statistic from this model. False discovery rate (FDR) was calculated using Benjamini-Hochberg (BH) method. We used these cis-eQTLs for the trans-eQTL analysis for the TSN edge replication described below.

Trans-eQTLs from TSNs

We computed trans eQTLs in two ways. First, we found the best cis-associated variant per gene (smallest p-value, from the cis-eQTLs described in the previous paragraph) in that tissue, if one existed, and measured association between that variant and every neighbor of that gene in the TSN using the linear model of Matrix-eQTL (Shabalin, 2012). Second, we measured association between all variants within 20 kb of a gene's TSS and TES with each neighbor in the network using the linear model of Matrix-eQTL (Shabalin, 2012). In both approaches, we controlled for the first three genotype PCs, sex, and platform, and used BH FDR ≤ 0.2 for multiple testing correction.

Trans-splicing QTLs from TWNs

We computed trans-splicing QTLs using two approaches. In the first approach, we used the best cis-associated variant per gene (smallest p-value) located within 1 Mb from the transcription start site (TSS) of the gene (The GTEx Consortium, 2017). Then for every TE node connected with an IR node in the network, we measured association between the gene's best cis-associated variant and all the isoform ratio neighbors using the linear model of Matrix-eQTL (Shabalin, 2012), controlling for the first three genotype PCs and genotype platform. We corrected for false discovery (BH FDR ≤ 0.05). In the second approach, for each of the top 500 TE-IR hubs, we took all variants within 20 kb of the TSS and tested their association with isoforms located on a different chromosome and connected with the TE hub using Matrix-eQTL. Here, we used FDR ≤ 0.2 for multiple tests correction.

3.3 Results

3.3.1 Reconstructing transcriptome-wide networks across human tissues

We aimed to capture a global view of transcription and splicing regulation across the transcriptome of diverse human tissues. Using RNA-seq data from the Genotype-Tissue Expression (GTEx) project (v6) (The GTEx Consortium, 2017), we reconstructed TWNs independently for each of the sixteen tissues with samples from at least 200 donors. Before applying our method, we corrected expression data from each tissue for known and unobserved

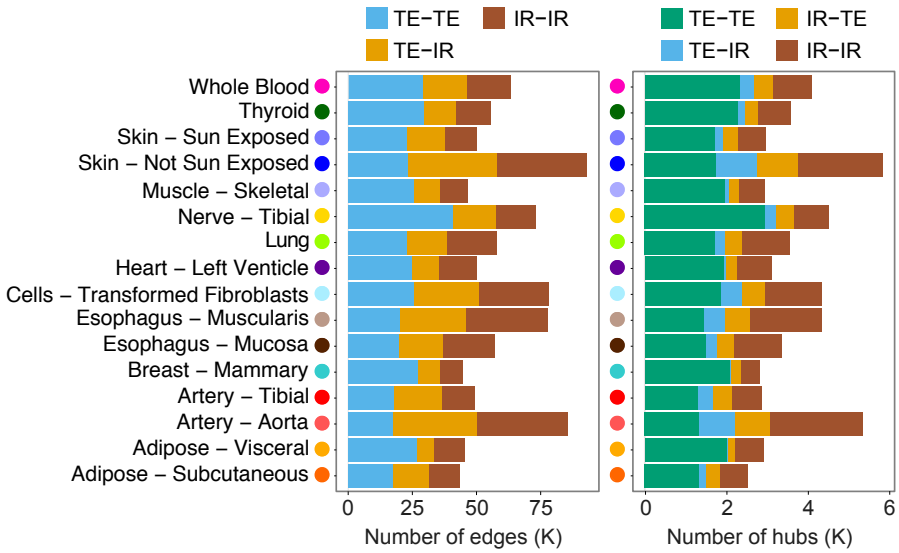

Figure 3.3: GTEx transcriptome-side networks summary. For each tissue, number of edges and number of hub nodes ($\geq$ 10 neighbors), segmented by the type of nodes connected by each edge.

confounding factors using HCP (Mostafavi et al., 2013), and all total expression and isoform ratio values were separately projected onto quantiles of a standard normal distribution. For computational tractability, we decided to use 6,000 genes and 9,000 isoforms for each tissue from available genes and isoforms. After applying graphical lasso, we excluded some edges from our networks for quality purpose and interpretability. Specifically, we excluded edges between nodes belonging to the same gene for downstream analysis. We also excluded edges between cross-mappable genes (Saha & Battle, 2018) and between genes with overlapping positions in the reference genome to avoid alignment and mapping artifacts. On average, each TWN contained 60,697 edges, with 24,527 between TE nodes (TE-TE), 18,539 between IR nodes (IR-IR), and 17,631 edges connecting TE and IR nodes (TE-IR) (Figure 3.3).

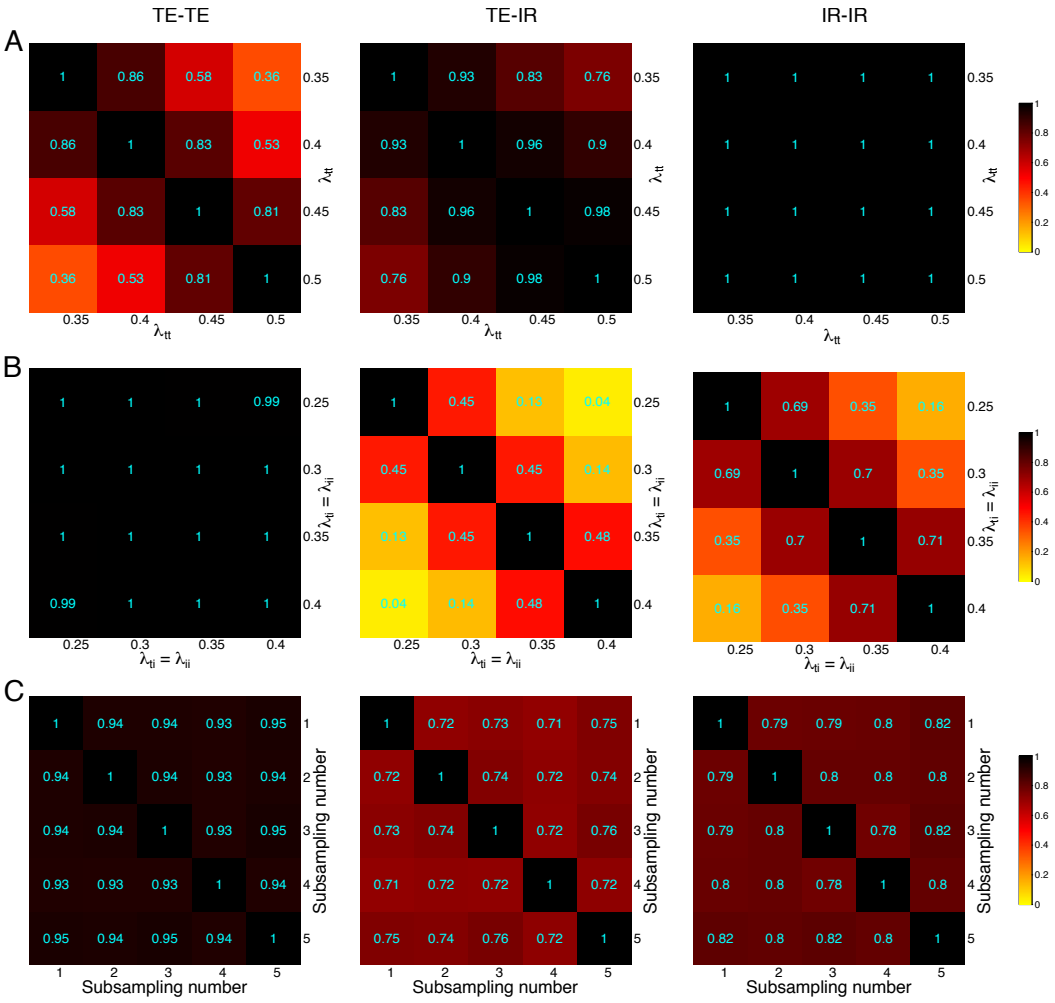

Figure 3.4: Robustness of TWN estimation for varying regularization parameters and sample size. We varied one of three variables at a time (regularization parameter λ_{tt}, regularization parameter λti, or sample size) keeping other variables the same as actually used, and re-estimated TWNs in *whole blood*. We then computed Tanimoto coefficients between edge weights of every pair of re-estimated TWNs categorized by type of edge: edge between two total expression nodes (TE-TE), between a total expression and an isoform node (TE-IR), and between two isoform nodes(IR-IR). A) Tanimoto coefficients for varying λ_{tt}. Tanimoto coefficients between the selected $\lambda_{tt}(0.4)$ and nearby choices $(0.35, 0.4)$ are very high in each category (0.86 and 0.83 for TE-TE, 0.93 and 0.96 for TE-IR, 1 and 1 for IR-IR, respectively). B) Tanimoto coefficients for varying $\lambda_{ti} = \lambda_{ii}$. Here, the selected $\lambda_{ti} = \lambda_{ii}$ was 0.25. C) Tanimoto coefficients for varying sample size. In each run, we randomly selected 90 % samples and re-estimated TWNs using the regularization parameters fixed to the same as actually used.

Reconstructing co-expression networks requires estimation of a large number of parameters (in our case, over 2×10^8) despite a small number of samples ($\leq$ 430); robustness and replicability of network edges are thus important considerations. The estimated networks were robust to change in regularization parameters and sample size in terms of similarity between networks measured by Tanimoto coefficient (Figure 3.4).

While there are not other large-scale RNA sequencing data sets for most GTEx tissue types, we replicated relationships identified by our GTEx whole blood TWN using an independent whole blood RNA-seq data set on 922 individuals of European ancestry from the Depression Genes and Networks study (DGN) (Battle et al., 2014; Mostafavi et al., 2014). First, we tested whether TE and IR nodes connected by an edge in the GTEx whole blood TWN were also correlated in DGN. For all edge types, we found that a higher fraction of gene pairs connected by an edge in the GTEx TWN were correlated in DGN compared to genes from random networks (84.7% versus 45.6%, 31.9% versus 5.9%, and 20.9% versus 2.6% for TE-TE, TE-IR and IR-IR edges, respectively; FDR $\leq$ 0.05; Figure 3.3B). Next, we reconstructed a TWN from DGN data over genes and isoforms common to both data sets. All pairs of nodes connected directly or indirectly in the GTEx whole blood TWN had significantly shorter network path distance in the DGN network compared to the distance in the same network with the node labels shuffled (Wilcoxon rank-sum test, $p \leq 2.2 \times 10^{-16}$; Figure 3.5). This provides replication in an independent sample for the same tissue, despite different alignment and isoform quantification pipelines between the two data sets.

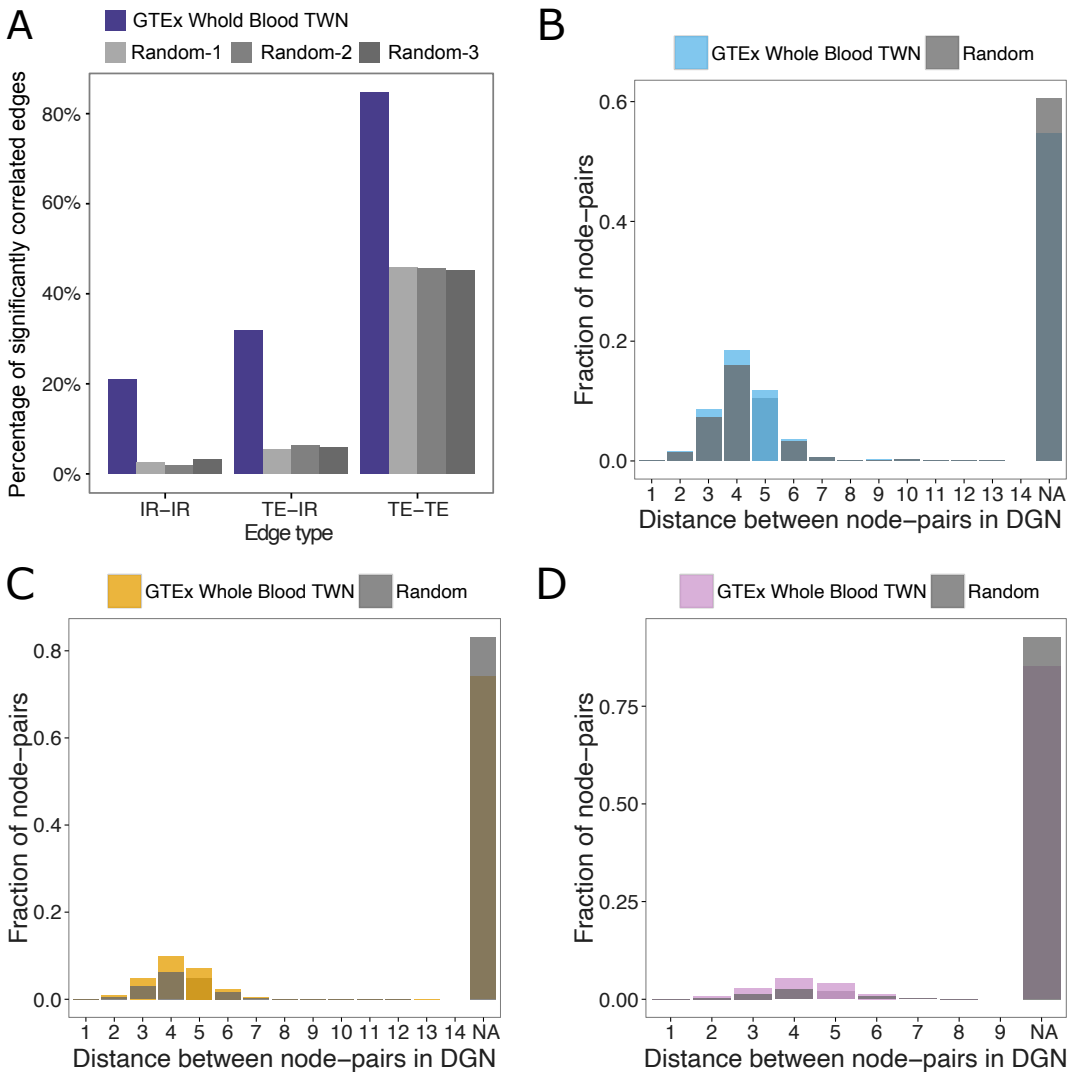

Figure 3.5: Replication of networks in an independent RNA-seq dataset. A) Percentage of edges from GTEx whole blood TWN edges that were significantly correlated in independent RNA-seq samples from DGN (Battle et al., 2014; Mostafavi et al., 2014). B-D) Fraction of connected node pairs from the GTEx whole blood TWN with a given distance between them in DGN TWN, categorized by node types: two total expression nodes (B), a total expression node and an isoform ratio node (C), and two isoform ratio nodes (D). In each plot, DGN networks are compared with random networks shown in gray.

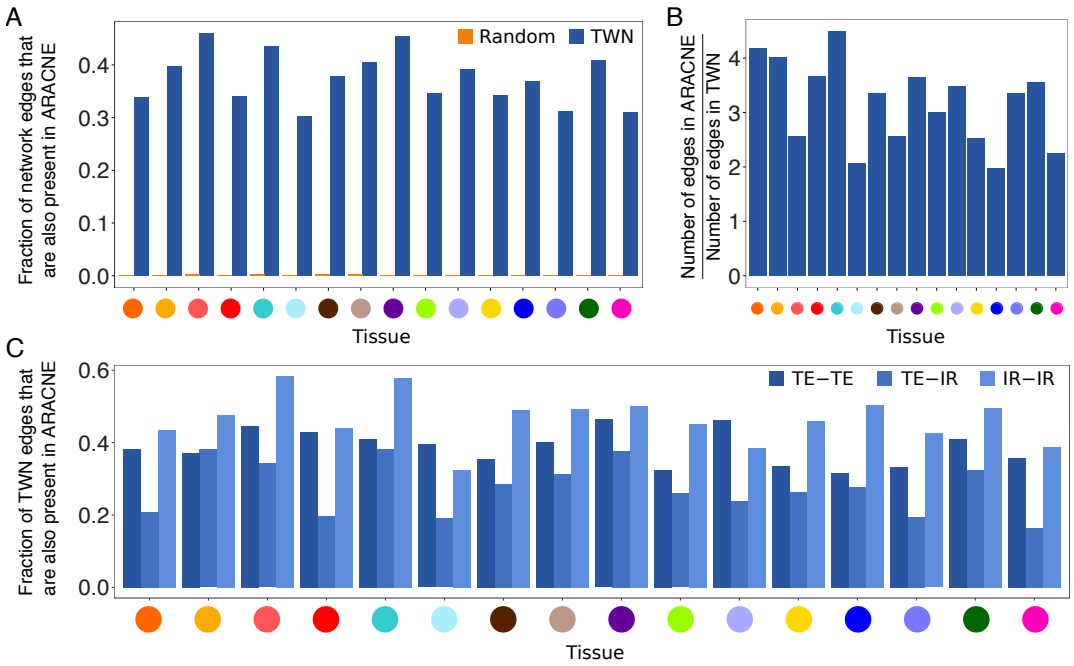

Figure 3.6: Replication of TWN using ARACNE. With the same data as used for TWNs, we reconstructed ARACNE networks from Spearman correlation based mutual information matrix using minet R package for 16 tissues. Following similar procedures as TWNs, we excluded edges between features of same gene, cross-mappable genes, and position-overlapped genes from downstream analysis. A) For TWNs and random networks, fraction of edges (y-axis) that were also present in ARACNE network in matched tissue (x-axis). A high fraction of TWN edges (30.42-46.34%, mean 37.72%), compared to random edges, were captured by ARACNE, demonstrating replication of TWN relationships using an independent method. B) Ratio of the number of edges in ARACNE network to the number of edges in TWN (y-axis) of the matched tissue (x-axis). On average, each ARACNE network had 3.17 times as many edges as the matched TWN indicating that TWN potentially captures direct relationships. C) Fraction of TWN edges (y-axis) that were also present in ARACNE network in matched tissue (x-axis), categorized by edge types. On average, 38.70%, 27.48%, and 46.43% of TE–TE, TE–IR, and IR–IR edges, respectively, were captured by ARACNE.

TWN relationships were also replicated by substituting a second gene regulatory network reconstruction method, ARACNE (Margolin et al., 2006) in place of graphical lasso, using the same overall framework and quantification of TE and IR levels in the GTEx data. ARACNE captured 37.73% of graphical lasso edges on average, compared to the expected proportion (0.15%) of edges captured at random (Figure 3.6), showing that the TWN signal is robust to choice of network estimation method.

3.3.2 TWN hubs are enriched for regulators of splicing

To characterize the TWNs, we focused on network hubs, as hub genes tend to be essential in biological mechanisms (Albert, 2005; Barabasi & Oltvai, 2004; Jeong et al., 2001). Unlike traditional networks, TWNs have four categories of hub genes that likely reflect different regulatory functions (Figure 3.1D). For instance, a hub arising from a total expression node connected to a large number of isoform ratio neighbors (TE-IR hub) may reflect a gene important in regulation of alternative splicing. We identified the top hub nodes by *degree centrality* – the number of edges per node – for each category. To avoid bias due to different numbers of isoforms per gene, we measured degree centrality of a node by the number of unique genes among neighboring nodes in each category. Based on a threshold of ten or more neighbors, TWNs had a mean of 1,853 “TE-TE" hub genes (total expression nodes connected to many total expression neighbors) and 325 “TE-IR" hub genes (total expression nodes connected to many isoform ratio neighbors) across tissues (Figure 3.3).

We investigated whether hub nodes with many IR neighbors were likely

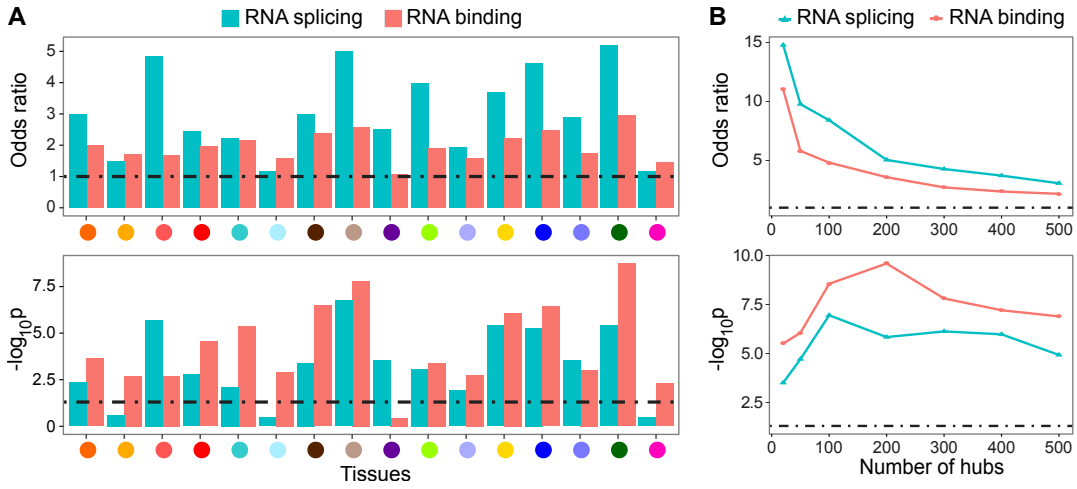

Figure 3.7: Enrichment of candidate splicing regulators among TWN hubs. A) In each TWN, the odds ratio and p-value of enrichment among the top 500 TE-IR hub genes for GO annotations reflect RNA binding and RNA splicing. B) Among consensus TE-IR hubs across all tissues, enrichment for GO annotations reflects RNA binding and RNA splicing functions.

to be regulators of alternative splicing. For each tissue, we evaluated the top TE-IR hubs for enrichment of Gene Ontology (GO) terms related to RNA splicing, and observed a significant abundance of known RNA splicing genes (annotated with GO:0008380) among the top TE-IR hubs. Indeed, 13 of 16 tissues (81.25%) showed significant enrichment of RNA splicing genes in the top 500 TE-IR hubs (significance assessed at Benjamini-Hochberg (BH) corrected $p \leq 0.05$; median across all tissues $p \leq 6.22 \times 10^{-4}$, Fisher's exact test), and every tissue had larger than unit odds ratio of RNA splicing genes among the top hubs (Figure 3.7A). Enrichment was robust to choice of hub degree threshold.

Next, we tested for enrichment of RNA-binding proteins, many of which are known to be important regulators of RNA splicing and processing (Chen & Manley, 2009; Wang & Burge, 2008; Witten & Ule, 2011). We found that RNA

binding genes (annotated with GO:0003723) were also significantly enriched, at BH corrected $p \leq 0.05$, among the top TE-IR hubs of every tissue except *heart – left ventricle* (median $p \leq 3.17 \times 10^{-4}$; Figure 3.7A). Across all GO terms, *splicing*, *RNA binding*, and *RNA processing* were consistently among the most enriched for TE-IR hubs across tissues (Tables A.1 and Table A.2). The replication network estimated from the DGN data also indicated relevant enrichment among TE-IR hubs (*RNA splicing*: $p \leq 1.07 \times 10^{-5}$, odds ratio 2.72; *RNA binding*: $p \leq 2.5 \times 10^{-11}$, odds ratio 2.37).

Many regulatory relationships are shared between tissues, and assessing hubs across all tissues jointly may improve robustness (Ballouz et al., 2015). Therefore, we identified TE-IR hubs shared across tissues (Table 3.1) using rank-product (Zhong et al., 2014). We first ranked hub genes according to the number of neighbors in each network. We then aggregated the ranks of those genes across all networks by computing the product of these ranks, and sorted genes to find the top TE-IR hubs (those with the largest number of neighbors in the most tissues). We observed much stronger enrichment for RNA splicing and RNA binding in the joint analysis than in individual tissues (Figure 3.7B).

Many of the top ranked TE-IR hubs shared across tissues are known to regulate splicing. *RBM14* (rank two), a RNA binding gene also known as *COAA*, interacts with a transcription regulator *TARBP2* to regulate splicing in a promoter-dependent manner (Auboeuf et al., 2004; Auboeuf et al., 2002). Another RNA binding gene *PPP1R10* (rank four) has been implicated in pre-mRNA splicing using mass spectrometry analysis (Du et al., 2014).

SRRM2 (rank eight) and *SRSF11* (rank nine) are also known splicing regulators (Blencowe et al., 2000; Chen & Manley, 2009; Wu et al., 2006; Zhang & Wu, 1996). For eleven of the top twenty cross-tissue TE-IR hubs, we found previous work supporting a role in the regulation of splicing (Table 3.1). These results suggest that TWN hubs are informative of splicing regulation, and uncharacterized TE-IR hub genes in a TWN are good candidates for regulatory effects on isoform abundance.

Co-regulation of expression and isoform ratios reflect biological pathways

Genes with similar function or that participate in the same pathway often have correlated patterns of gene expression (Hormozdiari et al., 2015; Khatri et al., 2012; Prieto et al., 2008; Roider et al., 2009). In the GTEx TWNs, we observed enrichment of edges between transcription factors and known target genes (Figure 3.8A). We also observed greater enrichment of closely connected genes for Reactome (Fabregat et al., 2016) and KEGG (Kanehisa et al., 2016) pathways as compared with permuted networks ($95 - 180$ Reactome, and $39 - 82$ KEGG pathways enriched per tissue at Bonferroni corrected $p \leq 0.05$; Wilcoxon rank-sum test; Figure 3.8B).

Patterns of correlation among relative isoform abundances are not well-studied, and it has not been established whether the regulation of splicing is coordinated across functionally related genes. Initial studies have identified such correlation in particular tissues (Iancu et al., 2015) and specific

Rank	Hub gene	#Tissues	Evidence
1	*TMEM160*	16	
2	*RBM14*	15	Nuclear receptor coactivator that interacts with *NCOA6* to regulate splicing in a promoter-dependent manner. (Auboeuf, Dowhan, Li, Larkin, Ko, Berget, & O'Malley, 2004; Auboeuf, Hönig, Berget, & O'Malley, 2002; Sui, Yang, Xiong, Zhang, Blanchard, Peiper, Dynan, Tuan, & Ko, 2007)
3	*ZMAT1*	16	
4	*PPP1R10*	15	Mass spectrometry analysis suggests its involvement in pre-mRNA splicing through interaction with *ZNF638* (Du et al., 2014)
5	*ODC1*	16	
6	*MGEA5*	16	
7	*KLHL9*	14	
8	*SRRM2*	15	Helps forming large splicing enhancing complexes (Chen & Manley, 2009). A mutation in *SRRM2* predisposes papillary thyroid carcinoma by changing alternative splicing (Tomsic et al., 2015).
9	*SRSF11*	14	A known serine/arginine-rich splicing factor (Wu, Kar, Kuo, Yu, & Havlioglu, 2006; Zhang & Wu, 1996)
10	*ZNF692*	15	
11	*ARGLU1*	16	Arginine/glutamate rich gene modulates splicing affecting neurodevelopmental defects (Magomedova et al., 2016).
12	*PPRC1*	16	Encodes protein similar to *PPARGC1* that regulates multiple splicing events (Martínez-Redondo et al., 2016).
13	*LUC7L3*	15	Regulates splice-site selection (Zhou et al., 2008) and affects cardiac sodium channel splicing regulation. (Gao et al., 2013)
14	*DUSP1*	16	
15	*FOSL2*	16	
16	*XPO1*	16	Interacts with *TBX3* (Kulisz & Simon, 2008) that regulates alternative splicing in vivo (mouse). (Kumar P. et al., 2014)
17	*PNISR*	15	Interacts with *PNN*, a suggested splicing regulator, and colocalizes with *SRrp300*, a known component of the splicing machinery (Zimowska et al., 2003).
18	*PNN*	12	Likely to be involved in RNA metabolism including splicing (Li et al., 2003)
19	*PTMS*	12	Involved in RNA synthesis processing (Vareli et al., 2000).
20	*CCDC85B*	15	

Table 3.1: Top 20 cross-tissue TE-IR hubs. (Rank) Rank-product rank of the gene; (#Tissues) number of tissues, out of 16, for which the hub gene (TE) has at least one IR neighbor.

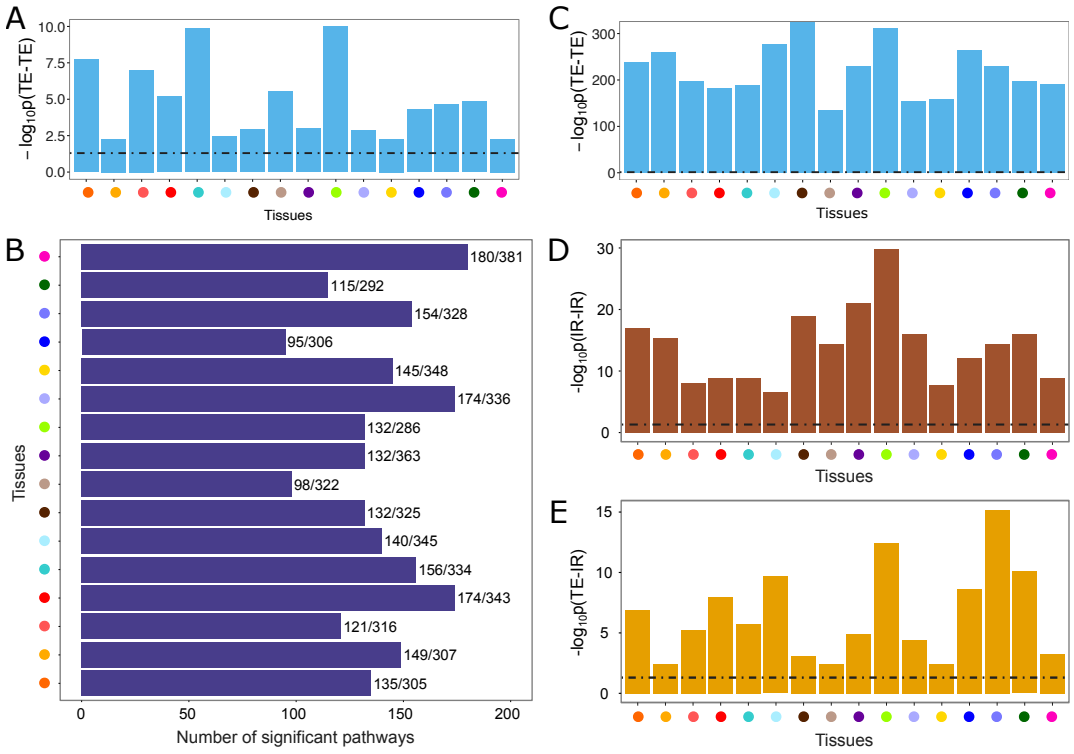

Figure 3.8: Pathway enrichment in TWNs. Tissue colors are matched with tissue names in Figure 3.3. A) Per-tissue, TF-Target enrichment among edges connected by two TE nodes. B) Per-tissue, the number of Reactome pathways enriched among connected components / total number of tested pathways for that tissue, considering only total expression nodes. C-E) Enrichment for shared Reactome pathway annotation among gene pairs connected by edges between two TE nodes (C), two IR nodes (D), and a TE and an IR node (E).

processes (Dai et al., 2012). To extend this, we evaluated each TWN for enrichment of edges between functionally related genes. For all 16 tissues, the TWNs demonstrated significant abundance of edges between isoform ratios of two distinct genes that participate in the same Reactome pathway (all tissues significant at BH corrected $p \leq 0.05$; median $p \leq 10^{-14}$; Figure 3.8D). Similarly, TE-IR edges were enriched for pairs of genes that participate in the same pathway (median $p \leq 10^{-5}$; Figure 3.8E). As expected, we also observed shared-pathway enrichment for nodes connected by TE-TE edges (Figure 3.8C). The patterns of functional enrichment were stronger among pairs of TE nodes, which may be due to more accurate quantification of total expression versus isoform ratios from RNA-seq data, functional annotations derived from gene expression studies, or tighter co-regulation of transcription than splicing among functionally related genes. Leveraging the co-regulation of splicing among functionally related genes, TWNs can be used to predict gene function (Warde-Farley et al., 2010) based on a more comprehensive understanding of co-regulation including regulation of splicing.

3.3.3 Comparison between TWNs reveals per-tissue hub genes

We evaluated the overall similarity of the TWNs between tissues. We tested concordance of hubs between each pair of tissues using Kendall's rank correlation computed over genes ordered by degree centrality (Figure 3.9). We observed greater than random levels of similarity between most tissues for all hub types (median $p \leq 1.0 \times 10^{-5}$ for each hub type), and functionally related tissues showed greater levels of similarity. For example, the two skin

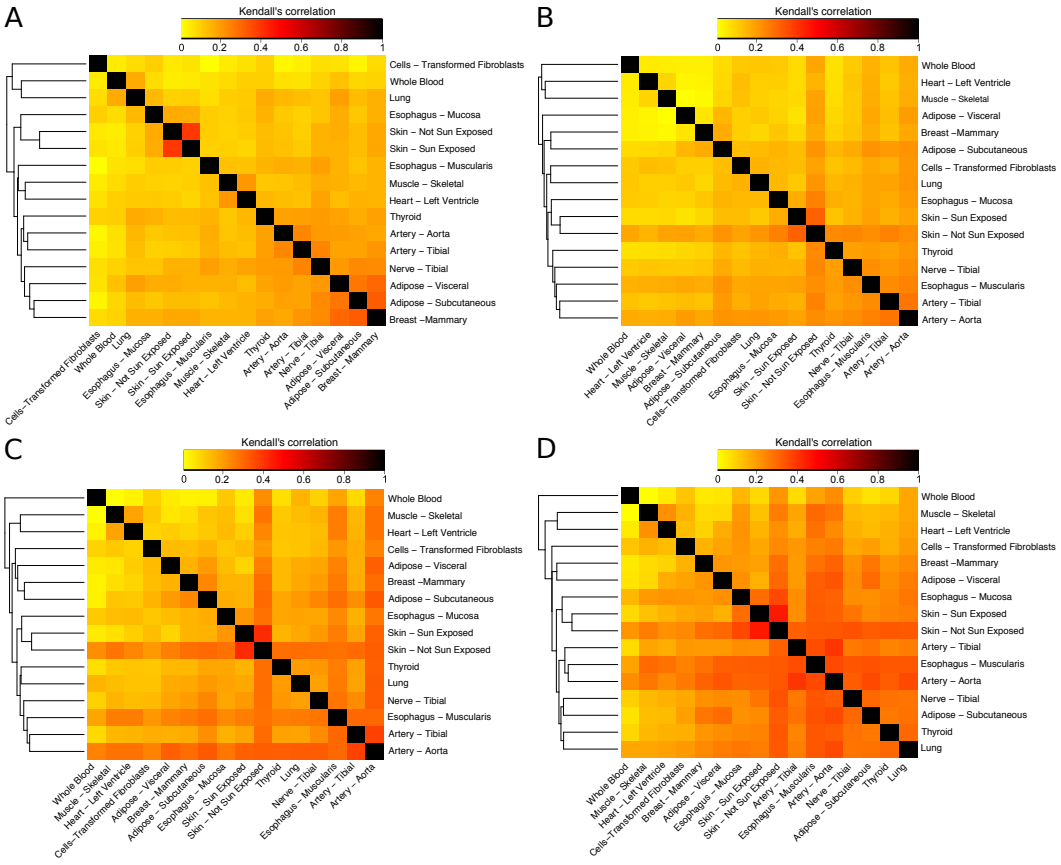

Figure 3.9: TWN Hub concordance. Heatmaps show Kendall's correlation coefficients between tissue pairs using ranking of TE-TE (A), TE-IR (B), IR-TE (C), and IR-IR(D) hubs. Tissue clustering dendrograms are shown at the left side of heatmaps.

tissues were grouped together for each hub type, and were found to be similar to *esophagus – mucosa*, which contains primarily epithelial tissue (Squier & Kremer, 2001). *Skeletal muscle* and *heart – left ventricle* grouped together, and *breast – mammary* was similar to the two adipose tissues, reflecting shared adipose cell type composition. While these results may be influenced by overlapping donors, they provide evidence that splicing is more similar in tissues with shared cell type compositions (Ong & Corces, 2011; Qian et al., 2005; The GTEx Consortium, 2017).

To identify candidate tissue-specific regulatory genes, we evaluated TE-IR hubs that had a high rank in related tissues, but a low rank among unrelated tissues. Several of the top ranked tissue-specific hubs were genes with evidence of known tissue-specific function or relevance. In the tissue group including *breast – mammary* and the two adipose tissues, the top tissue-specific TE-IR hub was *TTC36*, a gene highly expressed in breast cancer (Liu et al., 2008). The second ranked hub gene for the tissue group including *skeletal muscle* and *heart – left ventricle* was *LMOD2*, which was observed to be abundantly expressed in both tissues, and has been reported to regulate the thin filament length in muscles affecting cardiomyopathy in mice (Li et al., 2016b; Pappas et al., 2015).

We evaluated the tissue-specificity of our identified hub genes. To do this, we computed the fraction of top 100 TWN hubs of each tissue that did not appear in the list of top 500 TWN-hubs of any other tissue. We found that 8–43%, 11–39%, 0–24%, and 0–20% of our top 100 TE-TE, TE-IR, IR-TE, and IR-IR hubs, respectively, were uniquely identified in a single tissue. TE hubs (TE-TE

and TE-IR hubs) were more likely to be tissue-specific than matched IR hubs (IR-TE and IR-IR hubs; one-sided Wilcoxon signed rank test, $p \leq 4.13 \times 10^{-7}$). Tissue-specific hub proportions were not significantly different between TE-TE and TE-IR hubs (two-sided Wilcoxon signed rank test, $p \leq 0.52$). Many of the hub genes were differentially expressed across tissues (Table A.3).

An average of 69.87% of tissue-specific TWN edges connected nodes where at least one node was differentially expressed between the tissue of interest and all other tissues. For 6.9% of tissue-specific edges, at least one node was not included in a TWN for any other tissue because of low expression or other filters. However, for the remaining 23.22% of tissue-specific edges, both nodes were expressed in other tissues and included in other networks, so the tissue-specificity of edges is not exclusively due to expression levels.

3.3.4 Tissue-specific networks identify gene co-expression patterns unique to tissues

Transcriptome-wide networks (TWNs) reflect full gene network without distinguishing between tissue-shared and tissue-specific relationships. To directly assess the tissue-specificity of co-expression relationships, we built tissue-specific networks (TSNs) by considering all GTEx samples across 50 tissues *simultaneously*, decomposing the contributions to gene expression level variation into signals shared across tissues and those specific to single tissues. Taking genes that had non-zero counts in at least 10% samples in each and every tissue, we identified TSNs for 26 GTEx tissues. Here, we limited network nodes to total gene expression for simplicity. Across the 26 TSNs, the mean number of nodes (considering only genes with tissue-specific edges)

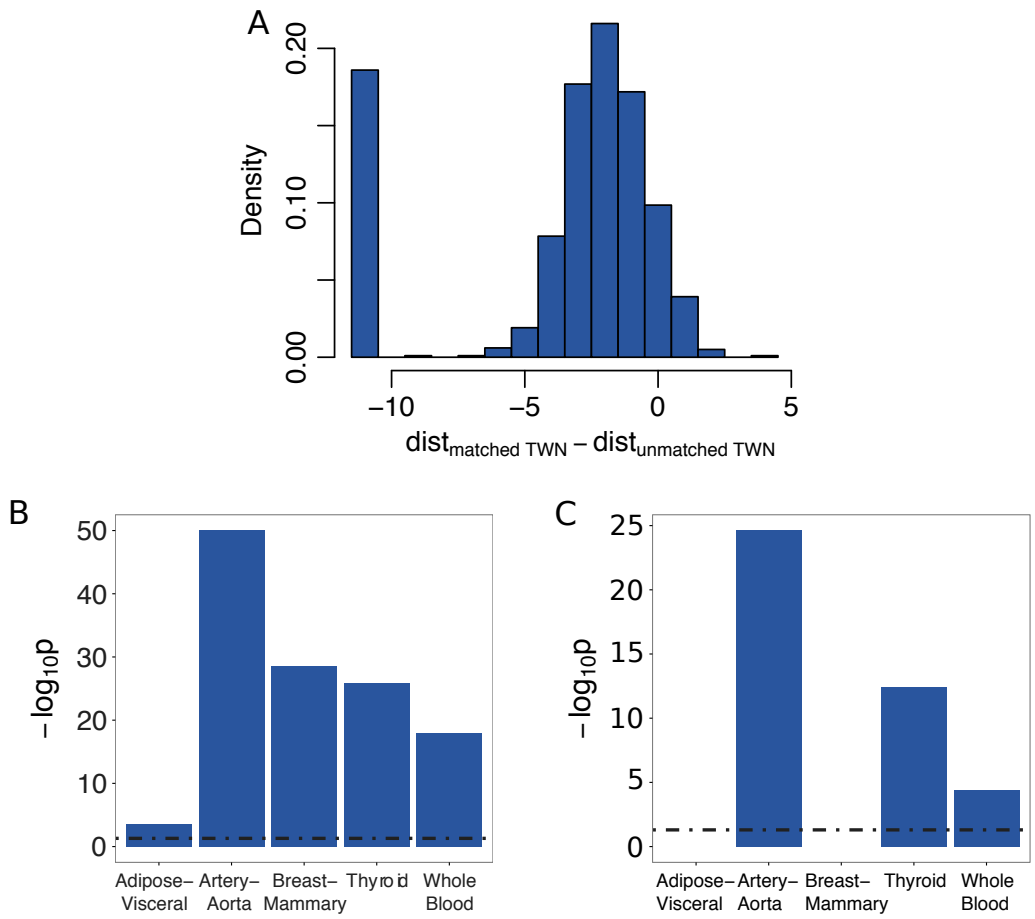

Figure 3.10: TSN edges were supported by TWNs and ARACNE networks. We selected edges from a TSN where both genes connected by an edge were jointly included in the matched TWN and at least another unmatched TWN. Then we tested if the path length between the nodes of a selected edge was significantly smaller in the matched TWN than in an unmatched TWN using one-sided Wilcoxon signed-rank test. A) Histogram of node distance differences for *artery-aorta*. As expected, most of the differences were negative, meaning that the nodes connected in *artery-aorta* TSN were closer to each other in *artery-aorta* TWN compared to other tissues' TWNs. B) BH corrected p-values for each tissue. Y-axis has been truncated to have a maximum value of 50. Note: a small number of edges (≤ 3) were selected for *lung*, *skeletal muscle*, and *esophagus–muscularis*, and p-values for these tissues were not reported. C) BH corrected p-values for the same test when ARACNE networks were used instead of TWNs.

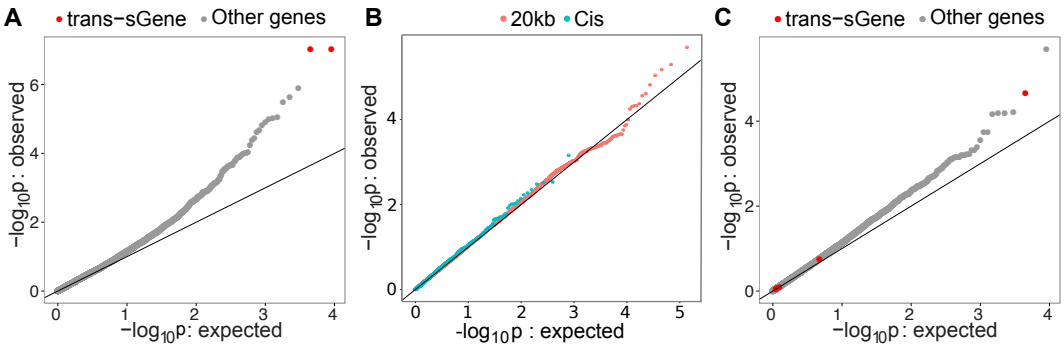

Figure 3.11: Association of local genetic variants with distant network neighbors. (A) Enrichment of association between rs113305055, a genetic variant near a cross-tissue TWN hub *TMEM160*, with all isoform ratios genome-wide in *artery – tibial*. (B) Enrichment of associations between local genetic variants (either the top cis-eVariant or any variant within 20 kb) of each gene, and network neighbors in the TSNs. (C) Enrichment of association between rs115419420, a genetic variant local to CRELD1, with all isoform ratios in skeletal muscle.

was 24, and the average number of edges was 107. As expected, TSNs contained a small subset of edges from full per-tissue TWNs, representing the co-expression components that are tissue-specific rather than shared. However, the signal in the TSNs is still reflected within their matched TWNs for the eight tissues where we reconstructed both networks (Figure 3.10).

3.3.5 Integration of networks with regulatory genetic variants

Both TWNs and TSNs were estimated using gene expression data alone. However, the GTEx v6 data also include genotype information for each donor. We intersected the edges detected by our networks with expression quantitative trait locus (eQTL) association statistics to replicate specific network edges through evidence of conditional associations with genetic variants across those edges and to increase power to detect long range (trans) effects of genetic variation on gene regulation.

First, we demonstrated that, for both TWNs and TSNs, there was enrichment for associations between the top cis-eVariant (the variant with lowest p-value per gene with a significant cis-eQTL) for each gene and the expression level or isoform ratio of its network neighbors based on QTL mapping in the corresponding tissue (Figure 3.11). This provides evidence of a causal relationship between connected genes. For TWNs, evaluating TE nodes with an IR neighbor, we found evidence for 61 trans (i.e., inter-chromosomal) associations and 86 intra-chromosomal associations tested between a cis-eVariant for the TE gene and the IR of the neighboring node (FDR ≤ 0.05). Our top two associations were between two variants, rs113305055 in *artery – tibial* and rs59153288 in *breast – mammary* (both near *TMEM160*), with isoform ratios of *CST3* ($p \leq 9.3 \times 10^{-8}$, and $p \leq 4.0 \times 10^{-7}$, respectively). *TMEM160* is the top cross-tissue hub in our TWNs with many IR neighbors (Table 3.1). Thus, we tested for association of these variants with all isoform ratios genome-wide and observed a substantial enrichment of low p-values in numerous tissues (Figure 3.11A; Figure A.1). In the TSNs, we identified five cis-eVariants across five tissues associated with six different trans-eGenes through six unique cis-eGene targets, one of which was intra-chromosomal (FDR ≤ 0.2; Table A.5). We also observed enrichment for low p-values over the tests corresponding to each network edge (Figure 3.11B).

We also performed a restricted test to identify novel trans-QTLs, without relying on the cis-eQTL signal from the same data, to avoid discoveries driven by potentially spurious correlations among expression levels. From the TWNs,

Variant	trans-eTranscript	trans-sGene	Local gene	P-value	FDR	Tissue
rs6122466	ENST00000496440.1	*CEP350*	*PPDP*	9.08×10^{-7}	0.09	Adipose – Visceral
rs397828484	ENST00000528430.1	*PPP1R16A*	*NMRK2*	1.66×10^{-6}	0.10	Muscle – Skeletal
rs7668429	ENST00000340875.5	*MEF2D*	*CLOCK*	4.81×10^{-6}	0.10	Muscle – Skeletal
rs7980880	ENST00000409273.1	*XIRP2*	*CALCOCO1*	9.91×10^{-6}	0.11	Muscle – Skeletal
rs56359342	ENST00000396435.3	*IQSEC2*	*CRAMP1L*	1.43×10^{-5}	0.14	Muscle – Skeletal
rs115419420	ENST00000531388.1	*CARNS1*	*CRELD1*	2.18×10^{-5}	0.19	Muscle – Skeletal

Table 3.2: Trans-sQTLs detected based on TWN hubs. (Variant) The most significant variant per trans-sGene listed; P-value and FDR for association between the variant and the trans-sGene listed; local gene target listed for reference.

we sought to identify trans-splicing QTLs (sQTLs) based on TE-IR hub genes, using the top 500 hubs by degree centrality. We tested every SNP within 20 kb of the TE hub-gene's transcription start site (TSS) for association with isoform ratios of each neighbor in the TWN. Using this approach, we identified 58 trans-sQTLs corresponding to six unique genes (sGenes) at FDR $\leq$ 0.2 (Table 3.2). For example, we identified a trans-sQTL association in *skeletal muscle* between rs115419420 and *CARNS1* ($p \leq 2.18 \times 10^{-5}$) that is supported by a cis association with the TE-IR hub *CRELD1*. This variant also showed enrichment for low p-values with numerous isoform ratios genome-wide (Figure 3.11C). In the TSNs, we identified 14 trans-eQTLs using variants within 20 kb of each gene and testing for association with the neighbors of those genes in the gene expression data of the same tissue (FDR $\leq$ 0.2; Table A.6). All of these associations were inter-chromosomal. Overall, we saw an enrichment of p-values for association between genetic variants local to a gene and the gene's neighbors in each network (Figure 3.11B).

3.4 Discussion

We reconstructed co-expression networks that capture novel regulatory relationships in diverse human tissues using large-scale RNA-seq data from the GTEx project. First, we specified an approach for integrating both total expression and relative isoform ratios in a single sparse Transcriptome-Wide Network (TWN). Splicing is a critical process in a number of tissue- and disease-specific processes and pathways (D'Souza et al., 1999; Ghigna et al., 2008; Glatz et al., 2006; Hutton et al., 1998), but, critically, isoform ratios have not been included in co-expression network analysis to allow the study of splicing regulation. We estimated TWNs from sixteen tissues and demonstrated that hubs in TWNs are strongly enriched for genes involved in RNA binding and RNA splicing. We found that, across tissues, the top hub genes with isoform ratio neighbors included many genes with known impact on splicing such as *RBM14*, a hub in all 16 tissues with TWNs. We identified a number of novel shared and tissue-specific candidate regulators of alternative splicing. While TWNs demonstrated clear enrichment for capturing desired regulatory relationships, care should be taken in interpreting individual edges and network relationships, as false positives may still arise due to confounding technical and biological factors, and from estimating large networks based on limited sample sizes. However, as more large-scale RNA-seq studies and better transcript quantification tools become available, TWNs will continue to be a useful and extensible framework for analyzing diverse types of regulatory relationships in disease, longitudinal, and context-specific studies.

Next, we estimated Tissue-Specific Networks (TSNs) for 26 single tissues.

These networks represent co-expression relationships unique to individual tissues.

Finally, we replicated edges in our networks using integration of genetic variation, and we identified 20 novel trans-QTLs affecting both expression and splicing. Together, our results provide the most comprehensive map of gene regulation, splicing, and co-expression in the largest set of tissues available to date. These networks will provide a basis for interpreting the transcriptome-wide effects of genetic variation, differential expression and splicing in complex disease, and the impact of diverse regulatory genes in the human genome.

3.5 Data and code availability

Software is publicly available on GitHub: https://github.com/battle-lab/twn_tsn. TWNs and TSNs for GTEx tissues are available at the GTEx portal (http://gtexportal.org).

Chapter 4

Inference and evaluation of co-expression networks

In the previous chapter, we used *graphical lasso* (Friedman et al., 2008) to study the regulation of transcription and splicing. In this chapter, we focus on a novel method to infer co-expression networks from gene expression profiles. The method is widely applicable to study the regulation of transcription. Our main contributions are listed below.

- We developed a novel method, SPICE, to reconstruct gene networks from gene expression profiles that prioritizes potential direct gene regulatory relationships.
- We also formulated a comprehensive set of evaluation metrics that use real biological data to compare network reconstruction methods.
- Using RNA-sequencing data of 49 tissues in humans from the Genotype-Tissue Expression (GTEx) consortium, we show that SPICE performs better than currently available methods in terms of most of our evaluation metrics.

This work has not yet been published.

4.1 Introduction

Gene co-expression networks (GCNs) capture the correlation between pairs of genes based on their expression profiles and can be used to infer gene regulatory relationships. However, the accuracy of GCNs remains low, especially for large complex networks (Chen & Mar, 2018; Guo et al., 2017; Marbach et al., 2012). Researchers need high-quality methods to estimate gene co-expression networks. Particularly, it is important to identify direct relationships between genes, as opposed to indirect associations via other genes, as direct relationships better illuminate functional mechanisms. To prioritize candidate direct regulatory relationships, we developed a novel method, SPICE, to reconstruct a gene network from gene expression profiles. SPICE utilizes maximum spanning trees to identify candidate direct edges in the network.

To evaluate network inference methods, researchers commonly use simulated data. However, biological data are generally more complex and noisier than simulated data. Methods that work well on simulated data may not perform similarly on real-world data (Marbach et al., 2012). A common approach in real-world network evaluation is to measure the area under the precision-recall curve (AUPR) of the learned network edges compared to a ground truth knowledge-base of known gene interactions. However, AUPR only captures a certain aspect of network fidelity to ground truth. Given our current knowledge of gene regulation is generally incomplete, noisy, and not

context-specific, network evaluation based on a single aspect of the current knowledge-base may not be reliable. To address this issue, we developed several evaluation metrics that capture diverse aspects of a gene network: i) does the network preserve the ranking of edges? ii) are the top hits precise? iii) how accurately does the network identify the hub genes? iv) do the connected genes share the same pathway? v) do the genes in a pathway cluster together? vi) how well can the network help detect trans-regulatory genetic variants? vii) how well can the network distinguish between direct and indirect interaction? These metrics enable us to benchmark network reconstruction methods comprehensively.

Using both simulated data and RNA-sequencing data of 49 tissues in humans from the GTEx consortium, we show that SPICE performs better than currently available methods in terms of most of our evaluation metrics. For example, based on protein-protein interactions from STRING (Szklarczyk et al., 2019), SPICE improves AUPR on average by 8.3%, rank preservation by 56%, and precision by 19.3%, relative to the best currently available methods. Based on the hallmark gene sets for humans in the molecular signature database (MSigDB) (Liberzon et al., 2005), SPICE improves pathway sharing by 14.7% relative to the best performance among currently available methods.

4.2 Methods

4.2.1 SPICE

SPICE is a computation framework to reconstruct a co-expression network from gene expression profiles. It is an ensemble learning algorithm that

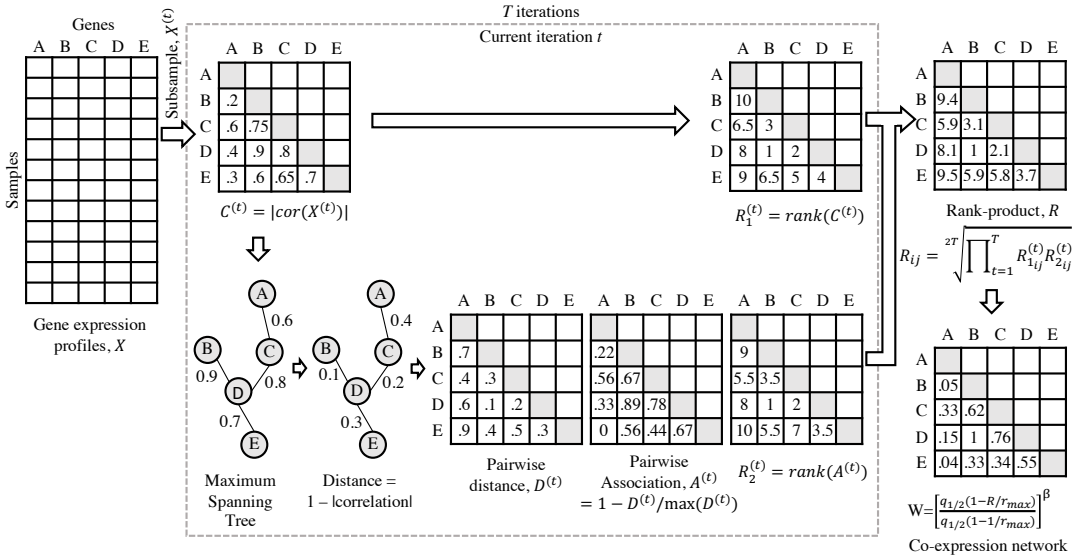

Figure 4.1: SPICE framework. SPICE is an ensemble learning algorithm to reconstruct a co-expression network (W) from gene expression profiles (X). In each iteration (t), it first computes gene-gene correlations ($C^{(t)}$) from sub-sampled expressions ($X^{(t)}$). It then generates two rank matrices ($R_1^{(t)}$ and $R_2^{(t)}$). $R_1^{(t)}$ directly ranks $C^{(t)}$, and $R_2^{(t)}$ ranks the same matrix via a maximum spanning tree, prioritizing potential direct regulatory relationships. Rankings from all iterations are aggregated using *rank-product*. Finally, SPICE assigns weights to each edge based on the rank-product to generate the co-expression network matrix (W).

aggregates rankings of edges between genes over multiple iterations. Given gene expression profiles of s samples across g genes ($X \in \mathbb{R}^{g \times n}$), SPICE estimates a co-expression matrix $W \in \mathbb{R}^{g \times g}$ representing relative rankings of edges (or interactions) between genes (Figure 4.1). Here, we assume that the expression values in X are processed, normalized, and already adjusted for confounding effects (Parsana et al., 2019; Stegle et al., 2012).

In each iteration, SPICE samples the given gene expression profiles, computes a gene-gene correlation matrix, and generates two ranking matrices. Formally, in t-th iteration ($1 \leq t \leq T$), we generate a sub-sampled gene expression matrix $X^{(t)}$ by randomly selecting a fraction of samples (default 80%) and genes (default 80%) from X using uniform sampling without replacement. Without loss of generality, let, $X^{(t)}$ preserves the order of the genes in X with NA's for genes not selected in t-th iteration. Then, we compute a pairwise correlation matrix $C^{(t)} \in \mathbb{R}^{g \times g}$, where (i, j)-th element $C_{ij}^{(t)} = |cor(X_i^{(t)}, X_j^{(t)})|$. Here, $X_i^{(t)}$ represents the expression of i-th gene across all selected samples. By default, we use Pearson correlation coefficient as *cor* function. Alternatively, normalized mutual information or Spearman correlation coefficient could be used. We assume *cor* function produces a symmetric correlation matrix where each element $\in [0, 1]$.

From the correlation matrix $C^{(t)}$, we generate two symmetric $g \times g$ ranking matrices, $R_1^{(t)}$ and $R_2^{(t)}$. Computation of $R_1^{(t)}$ is straight-forward: the lower triangle (excluding the diagonal) of $R_1^{(t)}$ contains the ranking of the lower triangle of $C^{(t)}$. The highest value is given a rank of 1, the 2nd highest value is given a rank of 2, and so on. In case of ties, the average ranking is used. The

diagonal elements are ignore (set to NA) and the off-diagonals are symmetric i.e., the (i, j)-th entry in the upper triangle of the ranking matrix is equal to the (j, i)-th element in the lower triangle.

$R_2^{(t)}$ is computed following a series of steps described below.

1. A maximum spanning tree (MST) is generated from the undirected graph corresponding to the correlation matrix $C^{(t)}$, where the weight of each edge between gene i and gene j is $C_{ij}^{(t)}$.

2. The weight of each edge in the MST is converted to a distance measure by subtracting the current weight from 1.

3. A pairwise distance matrix $D^{(t)}$ is computed where $D_{ij}^{(t)}$ = shortest distance between gene i and gene j in the MST. Assuming the edges in the MST represent potential direct regulatory relationships, the distance between directly connected genes would stay the same, but the distance between indirectly connected genes would increase.

4. $D^{(t)}$ is converted to a pairwise association matrix, $A^{(t)} = 1 - D^{(t)}/max(D^{(t)})$.

5. $R_2^{(t)}$ contains the ranking of $A^{(t)}$ computed in the same way as described in the previous paragraph.

Rankings across all iterations ($R_1^{(1)}, \ldots, R_1^{(T)}$ and $R_2^{(1)}, \ldots, R_2^{(T)}$) are aggregated using rank-product (R) – the geometric mean of all the rankings, where $R_{ij} = \sqrt[2T]{\prod_{t=1}^{T} R_{1_{ij}}^{(t)} R_{2_{ij}}^{(t)}}$. An edge between gene i and gene j will be ranked at the top (i.e., R_{ij} will be small) when it is ranked at the top in all or most iterations.

While R provides the ranking of all edges, it does not provide the relative weights. To provide relative weights for each edge, we convert the rankings to quantiles of a half-normal distribution (positive values only), standardize by dividing by the quantile of the maximum possible rank, $r_{max} = \binom{g}{2}$, and raise it to the power β. Formally, the weight of the edge between gene i and gene j, $W_{ij} = \left[\frac{q_{1/2}(1-R_{ij}/r_{max})}{q_{1/2}(1-1/r_{max})}\right]^{\beta}$, where $q_{1/2}(p)$ represents the quantile of an half-normal distribution at probability p, and β is selected in such a way that the network approximately follows a scale-free topology (similar to Weighted Gene Co-expression Network Analysis (WGCNA) (Langfelder & Horvath, 2008; Zhang & Horvath, 2005)).

The weight matrix W represents the co-expression network. We should note that the weights in W do not represent any physical or statistical property, rather they represent a relative ranking of edges.

4.2.2 Evaluation metrics

To evaluate co-expression network reconstruction methods, we developed a comprehensive set of evaluation metrics using real data (Figure 4.2). Contrary to evaluation using simulated data, our evaluation metrics provide an assessment of how well each method performs in practice. These metrics build on the current knowledge-base of transcription regulation and can be applied to diverse contexts to further our understanding.

Our evaluation framework uses four types of inputs: i) an estimated network (N_E), 2) a known network (N_K), 3) a pathway database (D), and 4) eQTL

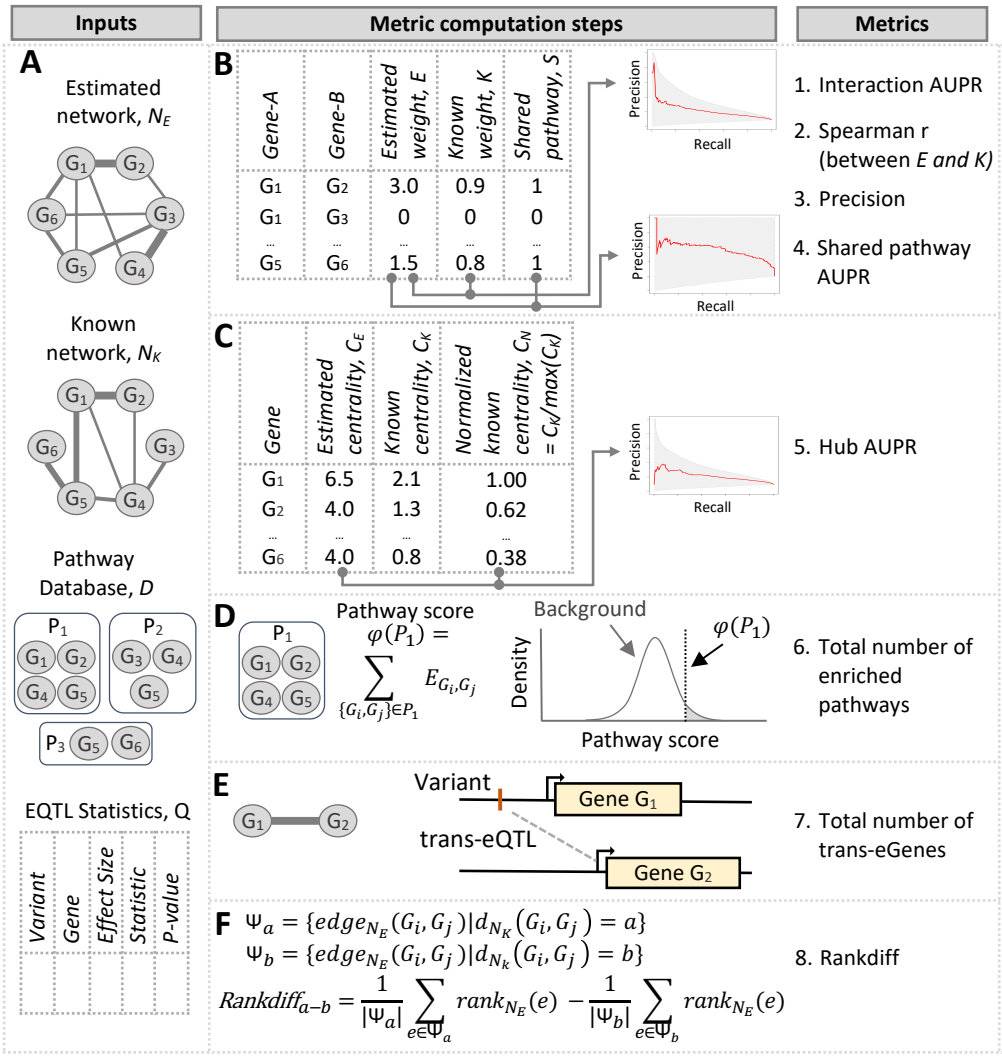
Inputs
Metric computation steps
Metrics
A
Estimated network, N_E
Known network, N_K
Pathway Database, D
EQTL Statistics, Q
Variant
Gene
Effect Size
Statistic
P-value
B
Gene-A
Gene-B
Estimated weight, E
Known weight, K
Shared pathway, S
G1 G2 3.0 0.9 1
G1 G3 0 0 0
G5 G6 1.5 0.8 1
Precision
Recall
1. Interaction AUPR
2. Spearman r (between E and K)
3. Precision
4. Shared pathway AUPR
C
Gene
Estimated centrality, C_E
Known centrality, C_K
Normalized known centrality, $C_N = C_K/max(C_K)$
G1 6.5 2.1 1.00
G2 4.0 1.3 0.62
G6 4.0 0.8 0.38
5. Hub AUPR
D
Pathway score
$\varphi(P_1) = \sum_{\{G_i,G_j\}\in P_1} E_{G_i,G_j}$
Background
Density
Pathway score
6. Total number of enriched pathways
E
Variant
Gene G1
trans-eQTL
Gene G2
7. Total number of trans-eGenes
F
$\Psi_a = \{edge_{N_E}(G_i,G_j)|d_{N_K}(G_i,G_j) = a\}$
$\Psi_b = \{edge_{N_E}(G_i,G_j)|d_{N_K}(G_i,G_j) = b\}$
$Rankdiff_{a-b} = \frac{1}{|\Psi_a|}\sum_{e\in\Psi_a} rank_{N_E}(e) - \frac{1}{|\Psi_b|}\sum_{e\in\Psi_b} rank_{N_E}(e)$
8. Rankdiff

Figure 4.2: Evaluation metric computation framework. A) The framework requires four types of inputs: an estimated network (N_E), a known network (N_K), a pathway database D, and eQTL statistics (Q). B) Each row in the table contains information about an edge (interaction) between two genes: the weight of the edge in N_E (E), the weight of the edge in N_K (K), and whether both genes share some pathway in D (S). *Interaction AUPR* is the area under the precision-recall (PR) curve computed using K as the ground truth probability and E as the classification score. *Spearman r* between E and K evaluates rank preservation. *Precision* is defined as the fraction of high-confidence ($K \geq$ *threshold*) edges in n top-weighted edges. *Shared pathway AUPR* is the area under the PR curve using S as the ground truth and E as the classification score. C) Each row in the table represents the centrality of a gene. Estimated centrality (C_E) and known centrality (C_K) of a gene are defined as the sum of weights of edges connected with the gene in N_E and N_K, respectively. *Hub AUPR* is the area under the PR curve using normalized known centrality (C_N) as the ground truth probability and C_E as the classification score. D) P-value of each pathway in D is the probability that the observed *pathway score* is greater than or equal to that of a random gene set of the same size. We compute the total number of *enriched pathways* in D at FDR ≤ 0.05. E) For each top-weighted edge in N_E, we test the variants nearby one gene for trans-association with the other gene and compute the total number of *trans-eGenes* using eQTL statistics Q. F) *Rankdiff* is the difference between average ranks of edges in N_E where the corresponding genes have certain distances in N_K.

statistics (Q) (Figure 4.2A). The estimated network (N_E) is a weighted undirected graph of genes, where the weight of each edge indicates the confidence of the edge i.e., how the method judges that the interaction between the genes is true.

The known network (N_K) is also a weighted undirected graph where the weights are within [0,1] range representing the probability that the corresponding genes interact with each other. Protein-protein interaction (PPI) networks such as STRING (Szklarczyk et al., 2019) and InWeb_IM (Li et al., 2016c) annotate each edge with such a score. Other binary PPI networks or gene-gene networks such as HuRI (Luck et al., 2020) can also be used with the probability score of either 0 or 1. STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) annotates a large number of PPIs with confidence scores based on experimental data, literature and database mining, and genomic context analysis. Importantly, STRING is well-maintained and regularly updated, making it a good candidate for network evaluation. We use STRING as the default choice of the known network.

Pathway databases are another important resource our framework uses. Conceptually, each pathway is a set of functionally relevant genes. A pathway database contains information about many pathways (or gene sets). KEGG (Kanehisa et al., 2016), REACTOME (Fabregat et al., 2016), BIOCARTA, Gene Ontology (Ashburner et al., 2000), etc. are some of the popular pathway databases. Molecular signature database (MSigDB) (Liberzon et al., 2005) contains a collection of pathway databases. Notably, MSigDB curates a list of 50 hallmark gene sets from multiple sources using a combination of automated

approaches and expert curation. These gene sets with reduced noise and redundancy represent well-defined biological states or processes and display coherent expression. In this work, we use MSigDB Hallmark gene sets as the pathway database unless stated otherwise.

Each row in the eQTL statistics table (Q) contains information about the test between a variant and a gene. In this project, we use eQTL statistics computed using data from the GTEx consortium (The GTEx Consortium, 2020).

Our evaluation framework defines eight metrics based on the above inputs.

1. **Interaction AUPR.** Detecting gene-gene interactions is a primary target of a gene co-expression network. *Interaction AUPR* measures how well an estimated network (N_E) captures the known interactions in N_K. Considering each edge weight of N_K as the ground truth probability, we compute the weighted area under the precision-recall curve (AUPR) of the estimated edge weights of N_E, which we call *interaction AUPR* (Figure 4.2B). The weighted AUPR is computed using the *pr.curve()* function in the *PRROC* R package. For this metric and a couple of other metrics defined later, we used the precision-recall curve, not the receiver operating characteristic (ROC) curve, due to the imbalance observed in the ground truth corresponding to a small number of positive examples in contrast to the large number of negative examples. This metric is commonly used to evaluate gene regulatory network reconstruction methods (Chen & Mar, 2018; Guo et al., 2017).

2. **Spearman ρ.** Spearman ρ is defined as the Spearman rank correlation

coefficient between the estimated weights and the known weights of all possible edges (Figure 4.2B). It assesses whether the estimated network preserves the ranking of the known network.

3. **Precision.** Top predictions from a network are often used in downstream analysis or follow-up studies (Ideker et al., 2001). Thus it is important for a network reconstruction method that its top predictions are true. The fraction of known high-confidence edges in n top-weighted edges from the estimated network is defined as *precision*. In the case of STRING, we considered edges with a score of 0.7 or higher as known high-confidence edges. We chose n to be the total number of known high-confidence edges between the genes used to reconstruct the network.

4. **Hub AUPR.** Hub genes play an important role in gene regulation (Saha et al., 2017; Seo et al., 2009; Seoane et al., 2019). To assess how the estimated network (N_E) captures known hubs in N_K, we first compute the estimated centrality (C_E) and the known centrality (C_K) defined as the sum of weights of edges the gene is connected to in N_E and N_K, respectively (Figure 4.2C). Then we divide C_K by $max(C_K)$ to get the normalized known centrality (C_N). Using C_N as the probability that the hub is true, we compute *Hub AUPR* as the weighted area under the precision-recall curve of C_E.

5. **Shared pathway AUPR.** Genes in a biological pathway generally have similar functions or share transcriptional regulatory programs. For evaluation based on this characteristic, we first define a random variable S indicating whether both genes corresponding to a given edge are

present in some pathway in D (Figure 4.2B). Then, considering S as an indicator of ground truth, we compute the weighted AUPR of the estimated edge weights (E), which we call *shared pathway AUPR*.

6. **Pathway enrichment.** Genes in a pathway are expected to cluster together in a network because of their functional similarity. If gene pairs from a given pathway are closer in the estimated network (N_E) than expected by chance, we state that the pathway is *enriched* in the network. Formally, we define a score for each pathway as the sum of absolute weights of edges between all possible gene pairs in the pathway. To get the null distribution of scores, we create 10,000 random gene sets with the same number of randomly selected genes, compute their scores, and fit a normal distribution. The enrichment p-value of the pathway is computed as $1 - cdf(\text{pathway score})$, where cdf refers to the cumulative distribution function. We use the total number of enriched pathways in D at $FDR \leq 0.05$ as an evaluation metric. In this work, we computed the enrichment of a pathway only if at least 5 genes and at most 100 genes from the pathway were included in the input data for network reconstruction.

7. **Number of trans-eGenes.** Detecting trans (inter-chromosomal) expression quantitative trait loci (trans-eQTLs) is computationally challenging because of the high number of tests and the corresponding high multiple test burden. Given that many trans-eQTLs are mediated by cis-associations (The GTEx Consortium, 2017, 2020), we can use relationships in a gene co-expression network to select a small number of

variant-gene pairs to test and thus increase the statistical power to detect trans-eQTLs (Saha et al., 2017). For each of the 10,000 top-weighted edges in the estimated network, we test genetic variants located within 1 Mb of one gene to the expression of the other gene. The total number of unique trans-eGenes (genes with at least one trans-eQTL) detected following this approach (FDR ≤ 0.05) quantifies the network's ability to help detect trans-eQTLs. We used the linear model in matrix-eQTL to call the trans-eQTLs from GTEx data.

8. **Rank difference.** It is expected that the direct interactions would be ranked higher (lower absolute value) than indirect interactions. To measure this quality, we first compute the mean rank of edges in the estimated network (N_E), where each edge corresponds to two genes with a given geodesic distance in the known network (N_K). Let, $\bar{R}_1$ represents the mean rank of edges in N_E where the genes are directly connected in N_K, $\bar{R}_2$ represents edges where the genes are 2-hop away from each other in N_K, and so on. Then we compute the difference between the mean ranks of two given geodesic distances a and b as $Rankdiff_{a-b} = \bar{R}_a - \bar{R}_b$. As an illustrative examples, a high positive $Rankdiff_{2-1}$ would mean that the network well distinguishes the direct interactions between pairs of genes from indirect interactions via another gene.

4.2.3 Implementation of gene co-expression networks

We implemented SPICE and benchmarked its performance against popular network reconstruction frameworks

SPICE. We implemented the SPICE method as an R package. The code is publicly available on GitHub. We used the *spice()* function in the package with default configurations as described in section 4.2.1.

WGCNA. We reconstruct a signed weighted gene co-expression network (Zhang & Horvath, 2005) as $((1 + cor)/2)^{\beta}$ where cor represents Pearson correlation between all gene pairs, and β is selected to make the network approximately scale-free using the pickSoftThreshold() function in the WGCNA R package (Langfelder & Horvath, 2008). We select the smallest $\beta \in \{ 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20 \}$ with a scale-free topology fitting index (Zhang & Horvath, 2005) $R^2 \geq 0.8$. If no β has $R^2 \geq 0.8$, we select the β with the highest R^2.

Graphical Lasso. We reconstruct a graphical lasso network (Friedman et al., 2008) by estimating a sparse inverse covariance matrix with L1-regularization using the glasso() function in the glasso R package. We do not penalize the diagonals and use a maximum of 1000 iterations. To select the regularization parameter (ρ), we split the gene expression matrix into two parts: a training data set with randomly selected 80% samples, and a validation data set with the rest 20% samples. For each $\rho \in \{0.05, 0.1, 0.15, ..., 0.5\}$, we estimate the inverse covariance matrix using the training data set, and then compute the held-out likelihood using the validation data set. We select the regularization parameter ρ that produced the highest held-out likelihood and call the glasso() function using all the samples. In this chapter, the graphical lasso may be abbreviated as *GLasso*.

MRNET and MRNETB. We first build a mutual information matrix using the

build.mim() function with *estimator* = "*pearson*" in the *minet* R package, and then call the *mrnet*() and *mrnetb*() function in the same package using the mutual information matrix to get *MRNET* (Meyer et al., 2007) and *MRNETB* (Meyer et al., 2010) networks, respectively.

GENIE3. We infer the GENIE3 network (Huynh-Thu et al., 2010) using the *GENIE*3() function in *GENIE*3 R package. For each target gene, we use an ensemble of 1000 trees using the Extra-Trees method.

ARACNE. To get an ARACNE network (Margolin et al., 2006), we first build a mutual information matrix using the *build.mim*() function with *estimator* = "*pearson*" and *disc* = "*equalfreq*" in the *minet* R package, and then call the *aracne*() function with *eps* = 0.1 in the same package using the mutual information matrix.

CLR. We inferred a CLR network (Faith et al., 2007) by first building a mutual information matrix using the *build.mim*() function with *estimator* = "*pearson*" and *disc* = "*equalfreq*" in the *minet* R package, and then calling the *clr*() function with default parameters.

Partial Correlation. We calculate the pairwise partial correlation coefficient for each pair of genes given others using the *pcor*() function with method = "spearman" in the *ppcor* R package and use it as a weight of the edge between the genes.

Random network. For each pair of genes, we generate a random number from a standard normal distribution and use its absolute value as the weight of the edge between the genes. A random network serves as a baseline for network reconstruction methods.

4.2.4 Symmetric absolute weighted network

To ensure higher weights represent higher confidence, we take absolute edge weights for each tissue and treat both positive and negative associations equally. If any method (e.g., GENIE3) produces an asymmetric network, then we convert it to a symmetric network by taking the maximum absolute value of weights of edges in two directions between each pair of genes.

4.2.5 Simulation

We simulated 5 data sets from multivariate normal distributions using the *huge* R package. Each data set had 200 samples. Each data set had an underlying network of 1,000 nodes (genes) and approximately 5,000 edges selected at a probability of 1%.

4.2.6 GTEx (v8) data

We downloaded fully processed, filtered, and normalized gene expression matrices for 49 tissue from the GTEx portal (http://gtexportal.org, GTEx_Analysis_v8_eQTL_expression_matrices.tar). Full details of data collection and processing are available in a paper by The GTEx Consortium (The GTEx Consortium, 2020). Briefly, RNA sequencing was performed using the Illumina TruSeqTM RNA sample preparation protocol. Sequencing was performed using HiSeq 2000 or 2500 to generate 76bp paired-end reads (median coverage 83M total reads). Reads were aligned to the human reference genome GRCh38/hg38 (excluding ALT, HLA, and decoy contigs) with STAR v2.5.3a. Gene-level read counts and TPM values were produced with RNA-SeQC v1.1.9 using the

"-strictMode" flag in RNA-SeQC. For each tissue, i) genes with TPM > 0.1 in at least 20% of samples and read counts ≥ 6 in at least 20% of samples were selected, ii) gene expressions were normalized between samples using TMM (Robinson & Oshlack, 2010), and iii) expression values of each gene were normalized across samples using an inverse normal transform.

We also downloaded the covariates for each tissue from the GTEx portal (GTEx_Analysis_v8_eQTL_covariates.tar.gz): 5 genotyping PCs, a set of covariates identified using the Probabilistic Estimation of Expression Residuals (PEER) method (Stegle et al., 2012), sequencing platform, sequencing protocol, and sex. We regressed out all available covariates from expressions for each tissue.

In our analysis, we used protein-coding genes from chromosomes 1-22 and X with unique HGNC gene symbols according to GENCODE v26 annotations. If multiple genes overlapped in their co-ordinates, we selected one of them with the highest variance. Finally, we selected the most variable 5000 genes from each tissue for further analysis.

4.3 Results

4.3.1 Simulation study

To demonstrate that our method can capture known regulatory relationships, we applied SPICE and current popular network inference methods on five simulated data sets. Each data set was generated from a multivariate normal distribution with a known underlying network structure of 1,000 genes and approximately 5,000 edges. Every method performed better than the baseline

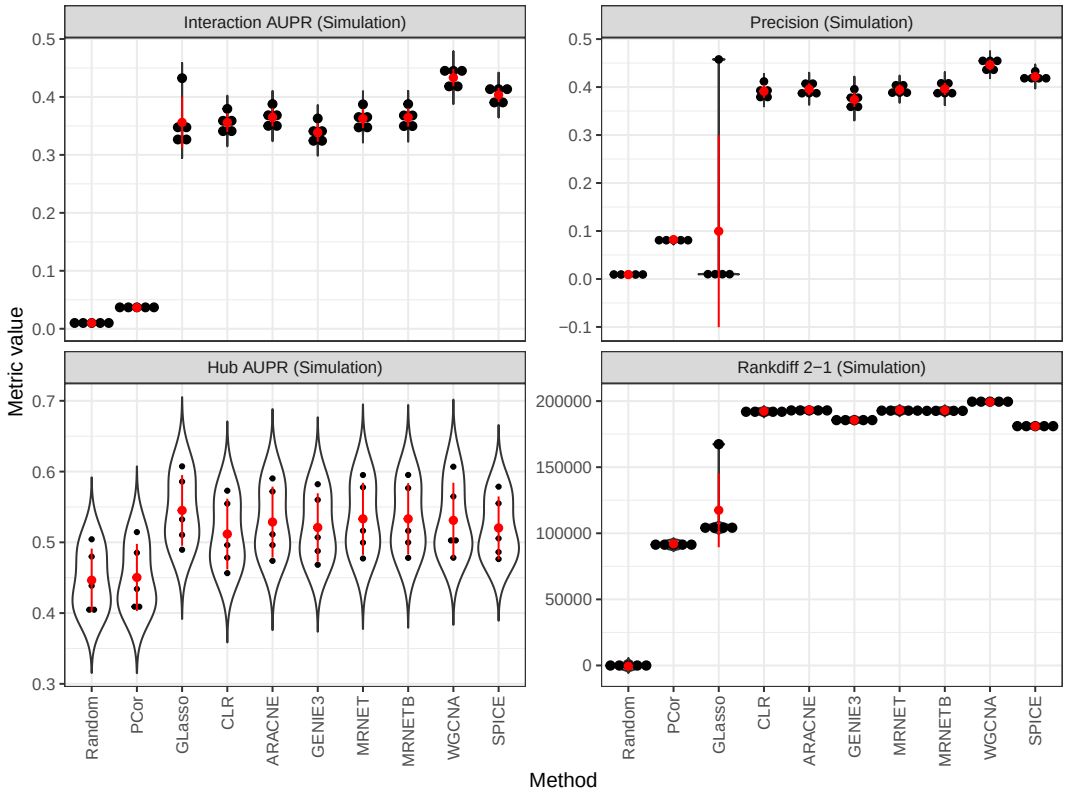

Figure 4.3: Network evaluation using simulated data. Each plot shows the performance of network inference methods (x-axis) in terms of an evaluation metric (y-axis, name at the top). Each black dot represents a method's performance using one of the five data sets. The red dot and the red bar represent the mean and the standard deviation of a method's performance across five data sets.

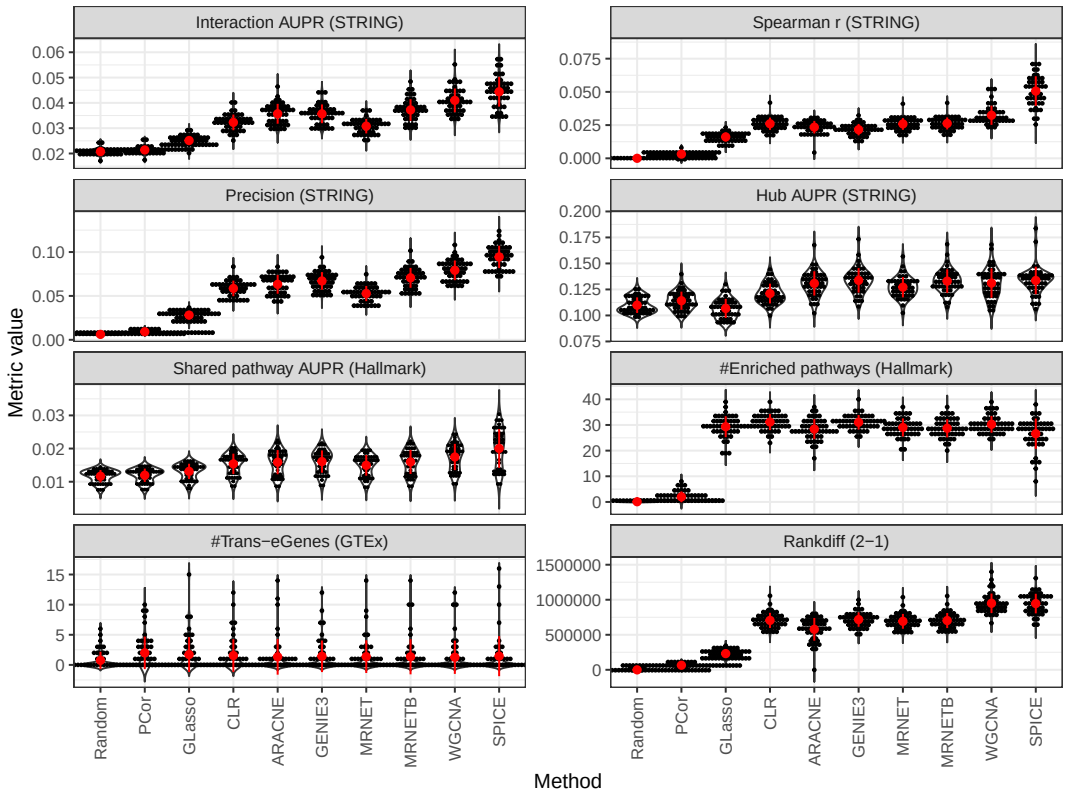

Figure 4.4: Evaluation of network inference methods. The performance of network reconstruction methods (x-axis) in terms of each evaluation metric (y-axis, name at the top). Each black dot represents a method's performance using one of the 49 tissues in GTEx. The red dot and the red bar represent the mean and the standard deviation of a method's performance across all tissues.

set by a random network (Figure 4.3). SPICE, along with WGCNA, MRNET, MRNETB, ARACNE, GENIE3, and CLR, was among the top performers to capture known regulatory relationships in terms of interaction AUPR, precision, hub AUPR, and rank difference. Notably, both interaction AUPR and precision were relatively low (0.35-0.45 and 0.38-0.45, respectively) even for the top-performing methods, highlighting the difficulty to estimate a large number of parameters from a relatively small number of samples.

4.3.2 Networks from diverse human tissues

Real biological networks are generally more complex and noisier than simulations. The underlying distribution of biological data is usually unknown. Gene collinearity, imprecise expression measurement, confounding factors, etc. further complicate the inference of biological networks.

To evaluate the performance of SPICE and other methods to estimate biological networks, we downloaded RNA-sequencing data of 49 tissues in humans from the Genotype-Tissue Expression (GTEx) consortium. After quality control, data processing, and confounding factor correction, each tissue had normalized gene expression of 5,000 highly variable autosomal protein-coding genes across 73-706 individuals.

Before applying each method to this data set, we optimized the parameters of each method using 1,500 genes across four tissues: Lung (515 samples), Heart – Left Ventricle (386 samples), Pancreas (305 samples), and Brain – Cerebellum (209 samples). We selected a parameter that resulted in a superior performance among possible options. For example, we used Extra-Trees (Geurts et al., 2006), instead of Random Forests, to build trees in GENIE3, because Extra-Trees generally produced superior performance in terms of our evaluation metrics. For the same reason, we estimated the mutual information in MRNET and MRNETB from a normal distribution instead of the empirical distribution.

Once the optimum parameters are selected, we applied SPICE and other methods to build gene co-expression networks for each tissue. Then, we computed our evaluation metrics for each network using the known network

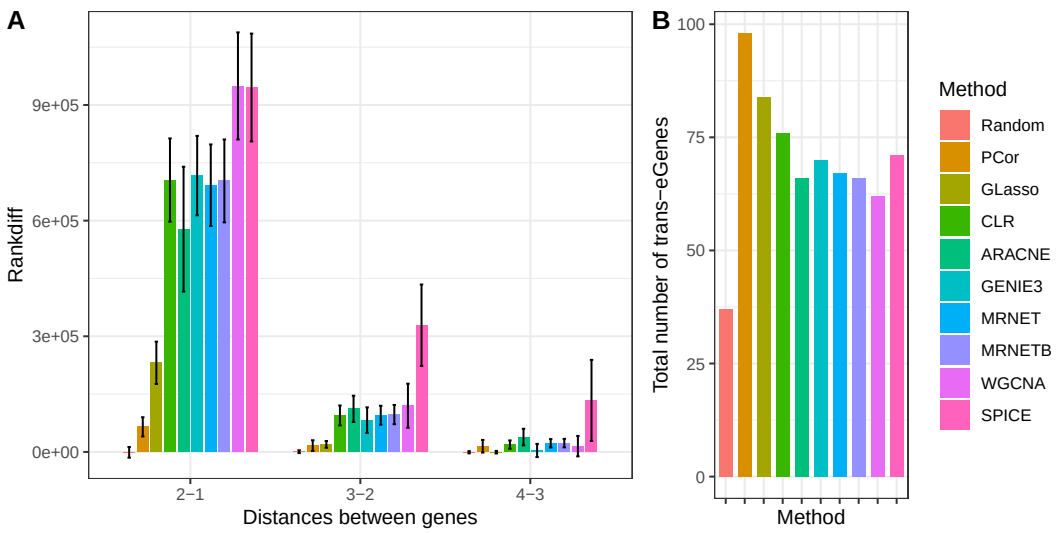

Figure 4.5: Rank difference (left) and the total number of trans-eGenes fromm all 49 tissues (right). '2-1', '3-2', and '4-3' in the left plot represents Rankdiff_{2-1}, Rankdiff_{3-2}, and Rankdiff_{4-3}, respectively.

from STRING, the pathway database from MSigDB Hallmark gene sets, and eQTL statistics from GTEx. Similar to the simulation analysis, all the methods performed, on average, better than the baseline set by the random network in terms of each metric (Figure 4.4). On average, SPICE was the best-performing method according to interaction AUPR, Spearman ρ, precision, and shared pathway AUPR. It improves interaction AUPR on average by 8.3%, rank preservation by 56%, precision by 19.3%, and shared pathway AUPR by 14.8% relative to the current best method WGCNA.

SPICE performed comparably to the top method in terms of hub AUPR, and rank difference between the 1st and 2nd-degree edges. Notably, SPICE had larger rank differences between higher-degree edges compared to other

methods (Figure 4.5). The performance was slightly low in terms of the number of enriched pathways (Figure 4.4). It may happen because the pathway enrichment, along with hub AUPR, depends on the absolute weights of edges, and SPICE computes the ranking of edges rather than absolute weights. For example, changing the power (β) to compute the final weight matrix, which does not change the edge ranking, may change the number of enriched pathways. In contrast, interaction AUPR, Spearman ρ, precision, and shared pathway AUPR depend on edge ranking, and SPICE performs well according to these metrics.

The number of trans-eGenes was small and highly-variable across tissues (Figure 4.4). It is not surprising owing to the difficulty of identifying trans-eQTLs. The full analysis of GTEx (v8) found only 121 protein-coding eGenes corresponding to 20 out of 49 tissues (The GTEx Consortium, 2020). So, instead of looking at the number of trans-eGenes from each tissue, we considered the total number of trans-eGenes from all tissues (Figure 4.5). Partial correlation detected the highest 98 trans-eGenes, and graphical lasso, which estimates an inverse covariance matrix proportional to partial correlation, found the second-highest number of 84 trans-eGenes. In comparison, SPICE found a total of 71 trans-eGenes comparable to other methods.

The evaluation of the network reconstruction methods was robust to the change in ground-truth resources. We tried alternative sources for the known interaction network (N_K): i) Experiment: each interaction score was derived only from experimental evidence as categorized by STRING (Szklarczyk et al., 2019), ii) InWeb_IM: The scored human protein-protein interaction network

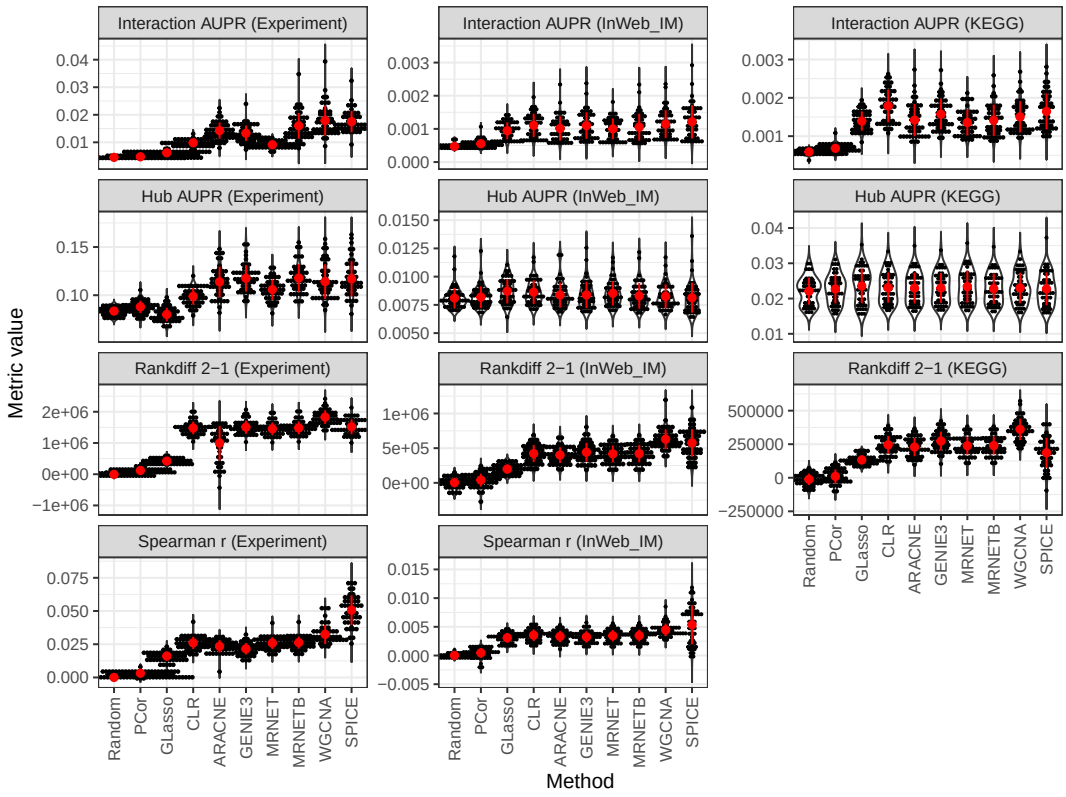

Figure 4.6: The evaluation framework is robust to change in the source of the known interaction network. The performance (y-axis) of the methods (x-axis) when the interaction scores originated from i) experimental evidence available in STRING (left column), ii) InWeb_IM (middle column), or iii) KEGG (right column).

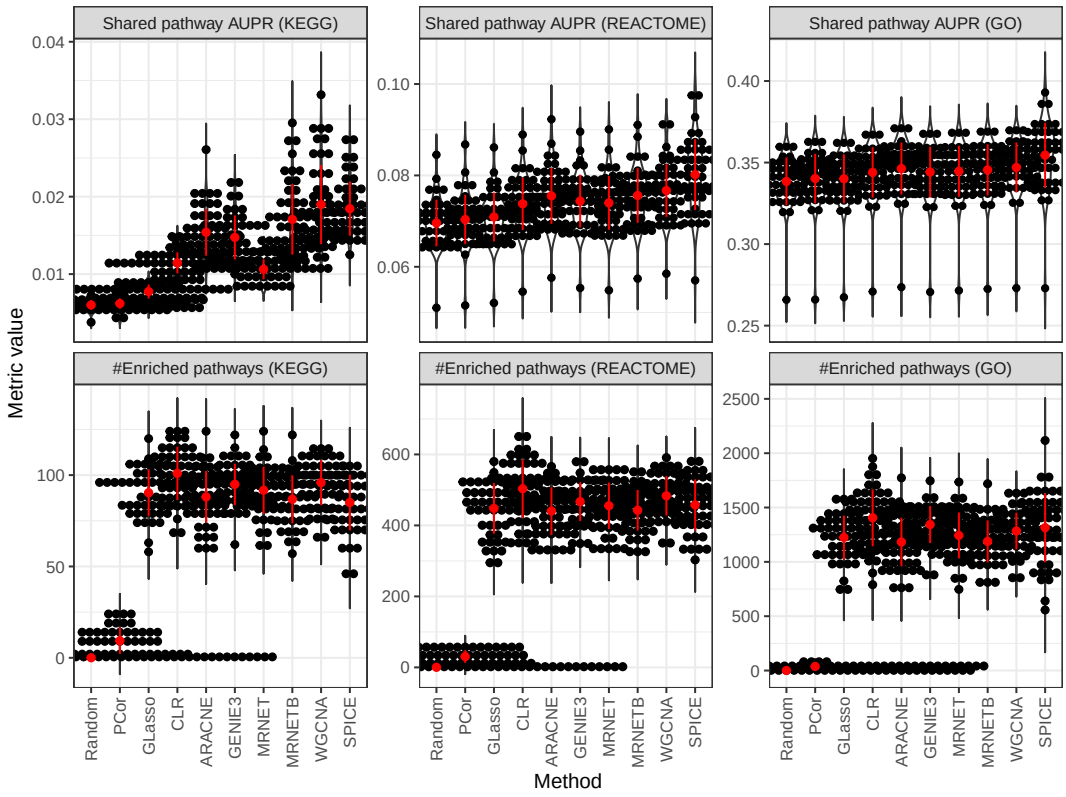

Figure 4.7: The evaluation framework is robust to change in the source of the pathway database. The performance (y-axis) of the methods (x-axis) when the pathway database originated from i) KEGG (left column), ii) REACTOME (right column), or iii) Gene Ontology (GO) (right column).

developed by Li et al., 2016c, and iii) KEGG: Binary gene-gene interaction annotated by KEGG (Kanehisa et al., 2016). Figure 4.6 show the corresponding performances. We also tried alternative sources for the pathway database (D): i) KEGG: gene sets derived from KEGG, ii) REACTOME: gene sets derived from the Reactome pathway database (Fabregat et al., 2016), and iii) GO: gene sets derived from gene ontology terms (Ashburner et al., 2000). Figure 4.7 show the corresponding performances.

Next, we compared the performances of SPICE and WGCNA – the current top-performing method in greater detail. We noticed that SPICE consistently performed better than WGCNA across almost all the tissues according to interaction AUPR, Spearman ρ, and precision. We noticed the opposite trend in terms of the number of enriched pathways. The relative performance of SPICE tends to improve with an increase in the sample size.

Next, to determine whether SPICE improved performances because of ranking computed using maximum spanning trees (MSTs) or because of multiple iterations, we ran SPICE excluding the ranking matrix $R_2^{(t)}$ computed using maximum spanning trees. This approach is essentially equivalent to running WGCNA for multiple iterations with subsamples and then aggregating the results using rank-product. On average, this approach performed equivalently to WGCNA and the standard SPICE performed better than this approach (Figure 4.9) suggesting that it is the ranking computed using MSTs that improves performances.

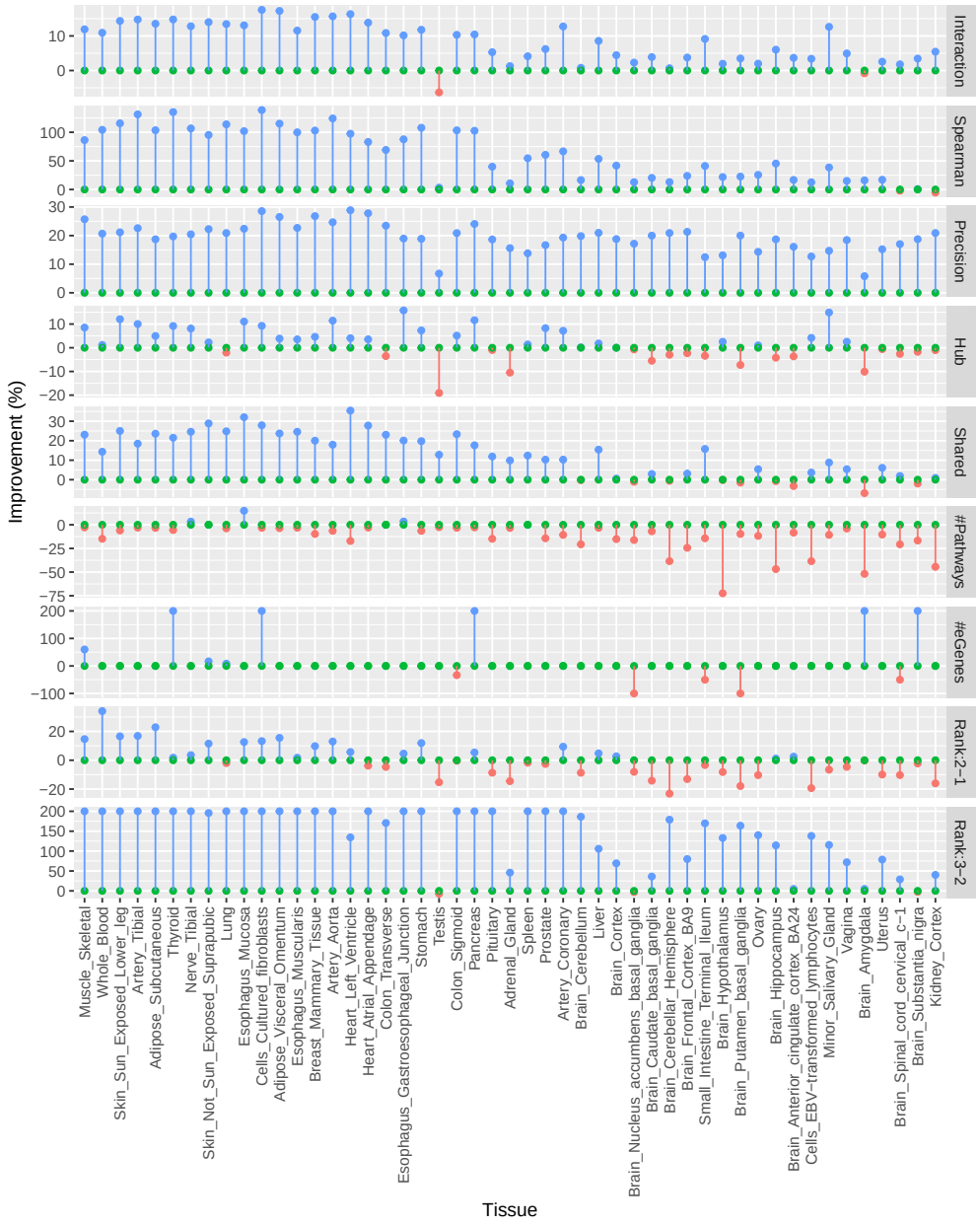

Figure 4.8: Comparison between SPICE and WGCNA. Improvement percentage (y-axis) achieved by SPICE relative to WGCNA in each tissue (x-axis), categorized by evaluation metrics (labels on the right). Improvement percentage has been clipped to a maximum of 200%. Tissues are sorted by decreasing samples size from left to right. *Interaction*: Interaction AUPR; *Spearman*: Spearman r; *Precision*; *Hub*: Hub AUPR; *Shared*: Shared pathway AUPR; *#Pathways*: Number of enriched pathways; *#eGenes*: Number of trans-eGenes; *Rank:2-1*: Rankdiff$_{2-1}$; *Rank:3-2*: Rankdiff$_{3-2}$.

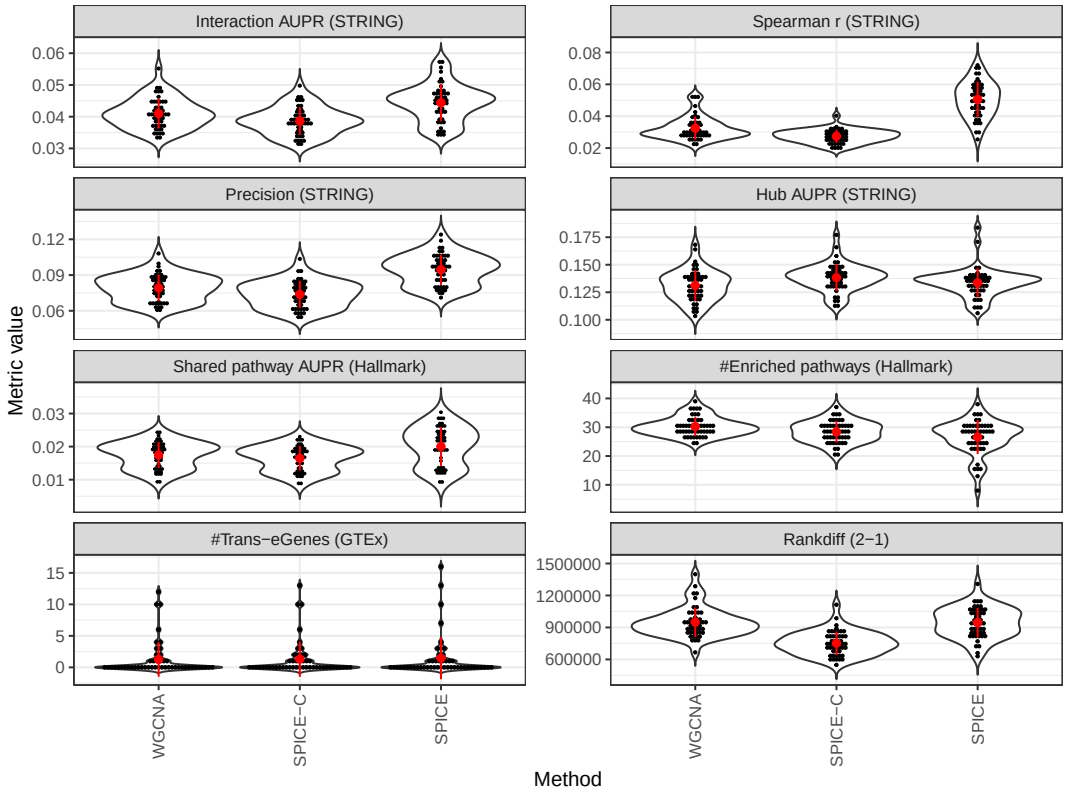

Figure 4.9: Ranking computed using maximum spanning trees improves performances. This plot shows performances (y-axis) of the WGCNA, SPICE-C, and SPICE (x-axis). SPICE-C refers to the approach where SPICE was run excluding the ranking matrix $R_2^{(t)}$ computed using maximum spanning trees.

Method	Muscle–Skeletal 706 samples	Lung 515 samples	Pancreas 305 samples	Brain–Cerebellum 209 samples
Random	10 sec (1)	5 sec (1)	5 sec (1)	5 sec (1)
CLR	21 sec (1)	18 sec (1)	12 sec (1)	10 sec (1)
WGCNA	32 sec (1)	38 sec (1)	25 sec (1)	22 sec (1)
ARACNE	5 min (1)	5 min (1)	4 min (1)	4 min (1)
PCor	18 min (1)	9 min (1)	9 min (1)	8 min (1)
MRNET	40 min (1)	19 min (1)	19 min (1)	19 min (1)
MRNETB	25 min (1)	25 min (1)	24 min (1)	24 min (1)
SPICE	31 min (1)	38 min (1)	35 min (1)	33 min (1)
GLasso	10.3 hr (1)	14.5 hr (1)	13.8 hr (1)	9.6 hr (1)
GENIE3	4.4 hr (6)	2.8 hr (6)	1.4 hr (6)	1 hr (6)

Table 4.1: Time (and cores) used to reconstruct gene co-expression network of 5,000 genes in a representative set of tissues.

4.3.3 Run time

SPICE ran faster than GENIE3 or Graphical lasso, slower than WGCNA, ARANCE, CLR, or Partial correlation, and comparable to MRNET or MERNETB. Here we report the time and the number of cores used to reconstruct networks from a representative set of tissues.

4.4 Discussion

Here, we presented a novel co-expression network inference method, SPICE, from gene expression profiles. SPICE uses maximum spanning trees to find and prioritize potential direct interactions between genes. We also presented eight metrics to evaluate co-expression networks using biological data. Evaluation using biological data is more appropriate to assess biological networks than that using simulated data. Because biology is generally more complex

and noisier than simulation. However, evaluation by simulations is useful to demonstrate that each method works when applied to a given data set with certain properties.

We first evaluated SPICE using simulated data to demonstrate that our method can capture relationships equivalently to the top-performing methods when applied to data generated from a multivariate normal distribution. Then using biological data from 49 human tissues in GTEx, we demonstrated that SPICE performs better than current popular methods in terms of four important metrics: interaction AUPR, spearman ρ, precision, and shared pathway AUPR. Relative to the current best-performing method WGCNA, SPICE improves these four metrics by 8.3%, 56%, 19.3%, and 14.7%, respectively, based on protein-protein interactions from STRING and hallmark gene sets from MSigDB. It also performs equivalently to other methods in terms of hub AUPR and rank difference.

SPICE performance is relatively low in terms of the total number of enriched pathways. This may happen because this metric uses absolute edge weights, but SPICE computes edge rankings. Though SPICE assigns weights based on a half-normal distribution, the weights merely represent rankings without any statistical or physical meaning. SPICE's low performance according to weight-based metrics and improved performance according to rank-based metrics together indicate that though SPICE may not be good at assigning weights, it produces a good ranking.

Interestingly, partial correlation and graphical lasso performed better than other methods in terms of the number of trans-eGenes. However, these two

methods generally performed poorly according to the rest of the metrics. SPICE performed equivalently to the other methods according to this metric.

Overall, SPICE and WGCNA performed the best and the second-best, respectively, among all methods. The ranking remained roughly the same when evaluated with alternative sources of ground truths, indicating the reliability of the evaluation framework. SPICE consistently performed better than WGCNA for almost all tissues in terms of interaction AUPR, spearman ρ, precision, and shared pathway AUPR, while the opposite happened in terms of the number of enriched pathways. Notably, WGCNA achieved a high rank-difference, $Rankdiff_{2-1}$, using both simulated and biological data. While a high rank-difference using simulated data without any collinear genes is expected, it is somewhat surprising that WGCNA achieved as high $Rankdiff_{2-1}$ as SPICE using STRING with possible collinear genes. Next, when we looked into per-tissue performances, we found that SPICE tends to improve with an increase in the number of samples. As sequencing technologies are improving rapidly and the cost of data collection is dropping, we think SPICE would be even more effective in the future.

4.5 Code availability

Codes to infer networks using SPICE and compute evaluation metrics are available as an R package on GitHub: https://github.com/alorchhota/spice.

Codes used for analysis presented in this chapter are also available on GitHub: https://github.com/alorchhota/spice_analysis.

Chapter 5

Conclusions

In this final chapter, we first summarise the findings in this thesis, and then discuss a few potential directions for future research.

5.1 Summary

In this thesis, we studied gene regulation in humans using two computational approaches: co-expression network and trans-eQTLs. We bridged a few gaps in current methods to detect regulatory patterns, applied current methods with appropriate modeling to study biological processes, and developed novel methods to discover gene regulatory relationships. The main contributions of this thesis are summarized below.

- In Chapter 2, we discussed potential artifacts in trans-eQTL and co-expression network studies arising from mapping errors due to sequence similarity between genes. We found that trans-eQTLs from human RNA-sequencing studies are strikingly affected by cross-mapping between genes. Over 75% of trans-eQTLs detected using a standard pipeline

were potentially false positives because of cross-mapping. The effect was not as striking in co-expression studies as in trans-eQTLs, but we still observed a higher than background fraction of cross-mappable genes in top correlated gene pairs. We demonstrated that replication studies cannot mitigate the concern of false positives, as cross-mapping is not a random error, but a systematic error. We proposed a metric named *cross-mappability* to quantify the potential for cross-mapping between genes. We applied this metric to identify trans-eQTLs from diverse human tissues using data from the Genotype-Tissue Expression (GTEx) consortium.

- In Chapter 3, we studied the joint regulation of transcription and alternative splicing in humans. We presented a framework called *Transcriptome-wide network (TWN)* to model both total gene expressions and relative isoform levels into a single sparse network. Using data from GTEx, we built and published transcriptome-wide networks for 16 human tissues. We demonstrated that the total expression hubs with multiple isoform neighbors in these networks are potential splicing regulators. We found literature evidence for about half of the top 20 cross-tissue hubs, indicating the novelty of the rest. We also analyzed the tissue-specificity of the regulatory relationships. Finally, we detected 20 genetic variants with a distant regulatory impact on transcription and splicing.

- In Chapter 4, we presented a novel network inference method named *SPICE* to study the regulation of transcription. *SPICE* uses maximum spanning trees to prioritize potential direct regulatory relationships. We

also formulated a comprehensive set of metrics to compare methods using biological data. These metrics establish a standard to evaluate biological networks. Using both simulated data and RNA-sequencing data of 49 tissues in humans from GTEx, we benchmarked SPICE against popular network inference methods. SPICE improves interaction identification on average by 8.3%, rank preservation by 56%, precision by 19.3%, and pathway sharing by 14.7% relative to the best currently available methods based on known protein-protein interactions from STRING and hallmark gene sets from the molecular signature database (MSigDB). Notably, SPICE improves relative performance with an increase in the number of samples.

5.2 Future directions

5.2.1 Benchmarking the effects of unmeasured confounding factor removal on network inference

Addressing batch effects and unmeasured confounding factors is challenging for network inference from transcriptomic data. A common approach is to regress out principal components (PCs) capturing non-random technical confounding factors. However, some researchers remain skeptical that the estimated PCs may contain biological signals and should not be regressed out. Even if one decides to remove PCs, it is unclear how many PCs should be regressed out. With the help of our suite of metrics to evaluate biological networks presented in Chapter 4, we are now prepared to evaluate the effects of removing PCs.

5.2.2 Incorporating prior knowledge into SPICE.

SPICE uses gene expression profiles to infer a network. However, because of the small number of samples relative to a large number of genes, SPICE may be under-powered to robustly detect regulatory relationships. We may increase the power by incorporating prior knowledge and transferring knowledge from related tissue or cell types.

5.2.3 Network inference for cell-type-specific gene regulation

The field of sequencing technology is very vibrant and rapidly moving. Current single-cell technology can perform sequencing from a single cell and quantify a diverse range of features. Single-cell data, as opposed to bulk data, allow us to study disease cells, rare cells (e.g., fetus, bone-marrow cells from a cancer patient), and micro-organisms independently, and thus enable a wide range of approachable questions. Understanding gene regulatory mechanisms for each cell type would enhance our knowledge and thus improve our capacity to control disease-causing genes. A challenge in current single-cell data is that the quantification is generally incomplete leading to extensive sparsity. Data is also noisy. Besides, multiple cells from the same individual do not satisfy the common assumption of independent and identically distributed (IID) random variables. By tackling these issues in a network inference model, we can infer cell-type-specific gene regulatory networks. Transfer learning by shared information between cell types could also improve network inference for each cell type.

5.2.4 Modeling genetic effects on pathways

Current eQTL studies commonly focus on the effects of a single genetic variant on a single gene. However, in complex diseases, a pathway (set of genes), rather than a single gene, is disrupted. Thus, pathway-regulating genetic variants might directly explain disease mechanisms. We may use dimensionality reduction techniques to represent pathway activities using a small number of features and find the genetic effect on these features. We can apply these methods to currently known pathways. We can also apply this method to find the genetic factors of the network modules susceptible to brain-diseases.

5.3 Concluding remarks

Overall, this thesis presented a range of computational methods to establish best practices in co-expression network and trans-eQTL analysis. Outcomes from these methods have the potential to explain biological processes and thereby improve human health.

References

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3891

Albert, R. (2005). Scale-free networks in cell biology. *Journal of Cellular Science*, *118*(Pt 21), 4947–4957.

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. https://doi.org/10.1186/gb-2010-11-10-r106

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29.

Auboeuf, D., Dowhan, D. H., Li, X., Larkin, K., Ko, L., Berget, S. M., & O'Malley, B. W. (2004). CoAA, a Nuclear Receptor Coactivator Protein at the Interface of Transcriptional Coactivation and RNA Splicing. *Molecular and Cellular Biology*, *24*(1), 442–453.

Auboeuf, D., Hönig, A., Berget, S. M., & O'Malley, B. W. (2002). Coordinate regulation of transcription and splicing by steroid receptor coregulators. *Science*, *298*(5592), 416–9.

Ballouz, S., Verleyen, W., & Gillis, J. (2015). Guidance for RNA-seq co-expression network construction and analysis: Safety in numbers. *Bioinformatics*, *31*(13), 2123–2130.

Barabasi, A. L., & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, *5*(2), 101–NIL.

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., Haudenschild, C. D., Beckman, K. B., Shi, J., Mei, R., Urban, A. E.,

Montgomery, S. B., Levinson, D. F., & Koller, D. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research*, *24*(1), 14–24.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M. C., Tassani, S., Piva, F., Perez-Amodio, S., Strippoli, P., & Canaider, S. (2013). An estimation of the number of cells in the human body. *Annals of Human Biology*, *40*, 463–471. https://doi.org/10.3109/03014460.2013.807878

Biotechnology, N. (2016). So long to the silos. *Nature Biotechnology*, *34*, 157. https://doi.org/10.1038/nbt.3544

Blencowe, B. J., Baurén, G, Eldridge, A. G., Issner, R, Nickerson, J. A., Rosonina, E, & Sharp, P. A. (2000). The SRm160/300 splicing coactivator subunits. *RNA*, *6*(1), 111–20.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Casper, J., Zweig, A. S., Villarreal, C., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., Lee, C. M., Lee, B. T., Karolchik, D., Hinrichs, A. S., Haeussler, M., Guruvadoo, L., Navarro Gonzalez, J., Gibson, D., Fiddes, I. T., Eisenhart, C., Diekhans, M., Clawson, H., … Kent, W. J. (2017). The UCSC Genome Browser database: 2018 update. *Nucleic Acids Research*, *46*(D1), D762–D769. https://doi.org/10.1093/nar/gkx1020

Chen, M., & Manley, J. L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nature Reviews Molecular Cell Biology*, *10*(11), 741–54.

Chen, S., & Mar, J. C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics*, *19*(1), 232. https://doi.org/10.1186/s12859-018-2217-z

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W., & Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, *10*(6), 563–569. https://doi.org/10.1038/nmeth.2474

Craig, N., Cohen-Fix, O., Green, R., Greider, C., Storz, G., & Wolberger, C. (2014). *Molecular biology: Principles of genome function*. Oxford University Press.

Dai, C., Li, W., Liu, J., & Zhou, X. J. (2012). Integrating many co-splicing networks to reconstruct splicing regulatory modules. *BMC Systems Biology*, *6*(Suppl 1), S17.

DeBoever, C., Ghia, E. M., Shepard, P. J., Rassenti, L., Barrett, C. L., Jepsen, K., Jamieson, C. H., Carson, D., Kipps, T. J., & Frazer, K. A. (2015). Transcriptome sequencing reveals potential mechanism of cryptic 3′ splice site selection in *SF3B1*-mutated cancers. *PLoS Computational Biology*, *11*(3), e1004105.

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., & Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, *25*(24), 3207–12. https://doi.org/10.1093/bioinformatics/btp579

DeLuca, D. S., Levin, J. Z., Sivachenko, A., Fennell, T., Nazaire, M.-D., Williams, C., Reich, M., Winckler, W., & Getz, G. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, *28*(11), 1530–2. https://doi.org/10.1093/bioinformatics/bts196

Derrien, T., Estellé, J., Sola, S. M., Knowles, D. G., Raineri, E., Guigó, R., Ribeca, P., Marco Sola, S., Knowles, D. G., Raineri, E., Guigó, R., & Ribeca, P. (2012). Fast computation and applications of genome mappability (C. A. Ouzounis, Ed.). *PLoS ONE*, *7*(1), e30377. https://doi.org/10.1371/journal.pone.0030377

DiLeo, M. V., Strahan, G. D., den Bakker, M., & Hoekenga, O. A. (2011). Weighted Correlation Network Analysis (WGCNA) Applied to the Tomato Fruit Metabolome (P. Csermely, Ed.). *PLoS ONE*, *6*(10), e26683. https://doi.org/10.1371/journal.pone.0026683

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21.

Du, C., Ma, X., Meruvu, S., Hugendubler, L., & Mueller, E. (2014). The adipogenic transcriptional cofactor ZNF638 interacts with splicing regulators and influences alternative splicing. *Journal of Lipid Research*, *55*(9), 1886–96.

Dutta, D., He, Y., Saha, A., Arvanitis, M., Battle, A., & Chatterjee, N. (2020). Novel Aggregative trans-eQTL Association Analysis of Known Genetic Variants Detect Trait-specific Target Gene-sets. *medRxiv*, 2020.09.29.20204388. https://doi.org/10.1101/2020.09.29.20204388

D'Souza, I., Poorkaj, P., Hong, M., Nochlin, D., Lee, V. M.-Y., Bird, T. D., & Schellenberg, G. D. (1999). Missense and silent tau gene mutations

cause frontotemporal dementia with parkinsonism-chromosome 17 type, by affecting multiple alternative RNA splicing regulatory elements. *Proceedings of the National Academy of Sciences*, *96*(10), 5598–5603.

Ensembl Genome Browser 102. (2020). Human assembly and gene annotation (GRCh38.p13). Retrieved February 10, 2021, from https://useast.ensembl.org/Homo_sapiens/Info/Annotation

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., et al. (2016). The Reactome pathway knowledgebase. *Nucleic Acids Research*, *44*(D1), D481–D487.

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., & Gardner, T. S. (2007). Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles (A. Levchenko, Ed.). *PLoS Biology*, *5*(1), e8. https://doi.org/10.1371/journal.pbio.0050008

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Gao, C., McDowell, I. C., Zhao, S., Brown, C. D., & Engelhardt, B. E. (2016). Context specific and differential gene co-expression networks via Bayesian biclustering. *PLoS Computational Biology*, *12*(7), e1004791.

Gao, G., Dudley, S. C., & Jr. (2013). RBM25/LUC7L3 function in cardiac sodium channel splicing regulation of human heart failure. *Trends in Cardiovascular Medicine*, *23*(1), 5–8.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Ghigna, C., Valacca, C., & Biamonti, G. (2008). Alternative splicing and tumor progression. *Current Genomics*, *9*(8), 556–570.

Gill, D., Arvanitis, M., Carter, P., Hernández Cordero, A. I., Jo, B., Karhunen, V., Larsson, S. C., Li, X., Lockhart, S. M., Mason, A., Pashos, E., Saha, A., Tan, V. Y., Zuber, V., Bossé, Y., Fahle, S., Hao, K., Jiang, T., Joubert, P., … Burgess, S. (2020). ACE inhibition and cardiometabolic risk factors, lung ACE2 and TMPRSS2 gene expression, and plasma ACE2 levels: a Mendelian randomization study. *Royal Society Open Science*, *7*(11), 200958. https://doi.org/10.1098/rsos.200958

Glatz, D. C., Rujescu, D., Tang, Y., Berendt, F. J., Hartmann, A. M., Faltraco, F., Rosenberg, C., Hulette, C., Jellinger, K., Hampel, H., et al. (2006). The alternative splicing of tau exon 10 and its regulatory proteins CLK2

and TRA2-BETA1 changes in sporadic Alzheimer's disease. *Journal of Neurochemistry*, *96*(3), 635–644.

Gong, J., Mei, S., Liu, C., Xiang, Y., Ye, Y., Zhang, Z., Feng, J., Liu, R., Diao, L., Guo, A.-Y., Miao, X., & Han, L. (2018). PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. *Nucleic Acids Research*, *46*(D1), D971–D976. https://doi.org/10.1093/nar/gkx861

Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., Zhang, R., Hartmann, B. M., Zaslavsky, E., Sealfon, S. C., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*, *47*(6), 569–576.

Greninger, A. L., Naccache, S. N., Federman, S., Yu, G., Mbala, P., Bres, V., Stryke, D., Bouquet, J., Somasekar, S., Linnen, J. M., Dodd, R., Mulembakani, P., Schneider, B. S., Muyembe-Tamfum, J.-J., Stramer, S. L., & Chiu, C. Y. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Medicine*, *7*(1), 99. https://doi.org/10.1186/s13073-015-0220-9

Grundberg, E., Small, K. S., Hedman, Å. K., Nica, A. C., Buil, A., Keildson, S., Bell, J. T., Yang, T.-P., Meduri, E., Barrett, A., Nisbett, J., Sekowska, M., Wilk, A., Shin, S.-Y., Glass, D., Travers, M., Min, J. L., Ring, S., Ho, K., … Consortium, T. M. T. H. E. R. M. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics*, *44*(10), 1084–1089. https://doi.org/10.1038/ng.2394

Guo, W., Calixto, C. P. G., Tzioutziou, N., Lin, P., Waugh, R., Brown, J. W. S., & Zhang, R. (2017). Evaluation and improvement of the regulatory inference for large co-expression networks with limited sample size. *BMC Systems Biology*, *11*(1), 62. https://doi.org/10.1186/s12918-017-0440-2

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., … Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research*, *22*(9), 1760–74. https://doi.org/10.1101/gr.135350.111

Hartl, C. L., Ramaswami, G., Pembroke, W., Saha, A., Parsana, P., Muller, S., Pintacuda, G., Lage, K., Battle, A., & Geschwind, D. H. (2020). The architecture of brain co-expression reveals the brain-wide basis of disease

susceptibility. *bioRxiv*, 2020.03.05.965749. https://doi.org/10.1101/2020.03.05.965749

Hawe, J., Saha, A., Waldenberger, M., Kunze, S., Wahl, S., Müller-Nurasyid, M., Prokisch, H., Grallert, H., Herder, C., Peters, A., Strauch, K., Theis, F., Gieger, C., Chambers, J., Battle, A., & Heinig, M. (2020). Network reconstruction for trans acting genetic loci using multi-omics data and prior information. *bioRxiv*, 2020.05.19.101592. https://doi.org/10.1101/2020.05.19.101592

Hormozdiari, F., Penn, O., Borenstein, E., & Eichler, E. E. (2015). The discovery of integrated gene networks for autism and related disorders. *Genome Research*, *25*, 142–154.

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, *5*(6).

Hutton, M., Lendon, C. L., Rizzu, P., Baker, M., Froelich, S., Houlden, H., Pickering-Brown, S., Chakraverty, S., Isaacs, A., Grover, A., et al. (1998). Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature*, *393*(6686), 702–705.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., Geurts, P., & Friedman, N. (2010). Inferring Regulatory Networks from Expression Data Using Tree-Based Methods (M. Isalan, Ed.). *PLoS ONE*, *5*(9), e12776. https://doi.org/10.1371/journal.pone.0012776

Iancu, O. D., Colville, A., Oberbeck, D., Darakjian, P., McWeeney, S. K., & Hitzemann, R. (2015). Cosplicing network analysis of mammalian brain RNA-Seq data utilizing WGCNA and Mantel correlations. *Frontiers in Genetics*, *6*.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., & Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, *292*(5518), 929–934. https://doi.org/10.1126/science.292.5518.929

Jeong, H, Mason, S. P., Barabási, a. L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, *411*(6833), 41–42.

Johnson, N. R., Yeoh, J. M., Coruh, C., & Axtell, M. J. (2016). Improved Placement of Multi-mapping Small RNAs. *G3 Genes | Genomes | Genetics*, *6*(7), 2103–11. https://doi.org/10.1534/g3.116.030452

Kahles, A., Behr, J., & Rätsch, G. (2016). MMR: a tool for read multi-mapper resolution. *Bioinformatics*, *32*(5), 770–772. https://doi.org/10.1093/bioinformatics/btv624

Kahles, A., Lehmann, K. V., Toussaint, N. C., Hüser, M., Stark, S. G., Sachsenberg, T., Stegle, O., Kohlbacher, O., Sander, C., Caesar-Johnson, S. J., Demchok, J. A., Felau, I., Kasapi, M., Ferguson, M. L., Hutter, C. M., Sofia, H. J., Tarnuzzer, R., Wang, Z., Yang, L., … Rätsch, G. (2018). Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell*, *34*(2), 211–224.e6. https://doi.org/10.1016/j.ccell.2018.07.001

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44*(D1), D457–D462.

Karimzadeh, M., Ernst, C., Kundaje, A., & Hoffman, M. M. (2018). Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Research*, *46*(20), e120–e120. https://doi.org/10.1093/nar/gky677

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, *8*(2).

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S. L., Morgan, M., Carey, V., Ellis, I., Green, A., Ali, S, Chin, S.-F., Palmieri, C, Caldas, C, Carroll, J., Wang, S, Feng, F., Chinnaiyan, A., Rossini, A., … Zhang, J. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), R36. https://doi.org/10.1186/gb-2013-14-4-r36

Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L., & Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, *20*(1), 278. https://doi.org/10.1186/s13059-019-1910-1

Kulisz, A., & Simon, H.-G. (2008). An evolutionarily conserved nuclear export signal facilitates cytoplasmic localization of the Tbx5 transcription factor. *Molecular and Cellular Biology*, *28*(5), 1553–64.

Kumar P., P., Franklin, S., Emechebe, U., Hu, H., Moore, B., Lehman, C., Yandell, M., & Moon, A. M. (2014). TBX3 Regulates Splicing In Vivo: A Novel Molecular Mechanism for Ulnar-Mammary Syndrome (G. S. Barsh, Ed.). *PLoS Genetics*, *10*(3), e1004247.

Lachmann, A., Xu, H., Krishnan, J., Berger, S. I., Mazloom, A. R., & Ma'ayan, A. (2010). ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, *26*(19), 2438–2444.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, *9*, 559. https://doi.org/10.1186/1471-2105-9-559

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, *10*(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25

Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., & Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome Research*, *14*(6), 1085–1094.

Lee, T. I., & Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell*, *152*(6), 1237–51. https://doi.org/10.1016/j.cell.2013.02.014

Lee, Y., Gamazon, E. R., Rebman, E., Lee, Y., Lee, S., Dolan, M. E., Cox, N. J., & Lussier, Y. A. (2012). Variants Affecting Exon Skipping Contribute to Complex Traits. *PLoS Genetics*, *8*(10).

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1), 1.

Li, C., Lin, R.-I., Lai, M.-C., Ouyang, P., & Tarn, W.-Y. (2003). Nuclear Pnn/DRS protein binds to spliced mRNPs and participates in mRNA processing and export via interaction with RNPS1. *Molecular and Cellular Biology*, *23*(20), 7363–76.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H.-D., Menon, R., Eksi, R., Guerler, A., Zhang, Y., Omenn, G. S., & Guan, Y. (2016a). A Network of Splice Isoforms for the Mouse. *Scientific Reports*, *6*, 24507.

Li, H.-D., Omenn, G. S., & Guan, Y. (2015). MIsoMine: a genome-scale high-resolution data portal of expression, function and networks at the splice isoform level in the mouse. *Database : The Journal of Biological Databases and Curation*, *2015*, bav045.

Li, S., Mo, K., Tian, H., Chu, C., Sun, S., Tian, L., Ding, S., Li, T.-R., Wu, X., Liu, F., Zhang, Z., Xu, T., & Sun, L. V. (2016b). Lmod2 piggyBac mutant mice exhibit dilated cardiomyopathy. *Cell & bioscience*, *6*, 38.

Li, T., Wernersson, R., Hansen, R. B., Horn, H., Mercer, J., Slodkowicz, G., Workman, C. T., Rigina, O., Rapacki, K., Stærfeldt, H. H., Brunak, S., Jensen, T. S., & Lage, K. (2016c). A scored human protein-protein interaction network to catalyze genomic interpretation. *Nature Methods*, *14*(1), 61–64. https://doi.org/10.1038/nmeth.4083

Li, W., Kang, S., Liu, C.-C., Zhang, S., Shi, Y., Liu, Y., & Zhou, X. J. (2014). High-resolution functional annotation of human transcriptome: predicting isoform functions by a novel multiple instance-based label propagation method. *Nucleic Acids Research*, *42*(6), e39.

Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., Gilad, Y., & Pritchard, J. K. (2016d). Rna splicing is a primary link between genetic variation and disease. *Science*, *352*(6285), 600–604.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2005). The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, *1*(6D1), 417–425. https://doi.org/10.1016/j.cels.2015.12.004

Lim, K. H., Ferraris, L., Filloux, M. E., Raphael, B. J., & Fairbrother, W. G. (2011). Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proceedings of the National Academy of Sciences of the United States of America*, *108(27)*, 11093–11098. https://doi.org/10.1073/pnas.1101135108

Liu, Q., Gao, J., Chen, X., Chen, Y., Chen, J., Wang, S., Liu, J., Liu, X., & Li, J. (2008). HBP21: a novel member of TPR motif family, as a potential chaperone of heat shock protein 70 in proliferative vitreoretinopathy (PVR) and breast cancer. *Molecular Biotechnology*, *40*(3), 231–240.

López-Bigas, N., Audit, B., Ouzounis, C., Parra, G., & Guigó, R. (2005). Are splicing mutations the most frequent cause of hereditary disease? *FEBS Letters*, *579*(9), 1900–1903.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Luck, K., Kim, D. K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F. J., Charloteaux, B., Choi, D., Coté, A. G., Daley, M., Deimling, S., Desbuleux, A., Dricot, A., Gebbia, M., Hardy, M. F., Kishore, N., … Calderwood, M. A. (2020). A reference map of the human binary protein interactome. *Nature*, *580*(7803), 402–408. https://doi.org/10.1038/s41586-020-2188-x

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

Magomedova, L., Tiefenbach, J., Zilberman, E., Voisin, V., Robitaille, M., Gueroussov, S., Irimia, M., Ray, D., Patel, R., Xu, C., Jeyasuria, P., Bader, G. D., Hughes, T. R., Krause, H., Blencowe, B. J., Angers, S., & Cummins, C. L. (2016). ARGLU1 is a Glucocorticoid Receptor Coactivator and Splicing Modulator Important in Stress Hormone Signaling and Brain Development. *bioRxiv*. https://doi.org/10.1101/069161

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., & Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, *9*(8), 796–804. https://doi.org/10.1038/nmeth.2016

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., & Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, *7 Suppl 1*, S7. https://doi.org/10.1186/1471-2105-7-S1-S7

Martínez-Redondo, V., Jannig, P. R., Correia, J. C., Ferreira, D. M. S., Cervenka, I., Lindvall, J. M., Sinha, I., Izadi, M., Pettersson-Klein, A. T., Agudelo, L. Z., Gimenez-Cassina, A., Brum, P. C., Dahlman-Wright, K., & Ruas, J. L. (2016). Peroxisome proliferator-activated receptor

gamma coactivator-1 alpha isoforms selectively regulate multiple splicing events on target genes. *Journal of Biological Chemistry*, *291*(29), 15169–15184.

Matlin, A. J., Clark, F., & Smith, C. W. J. (2005). Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology*, *6*(5), 386–98.

Mei, H., Xia, T., Feng, G., Zhu, J., Lin, S. M., & Qiu, Y. (2012). Opportunities in systems biology to discover mechanisms and repurpose drugs for CNS diseases. *Drug Discovery Today*, *17*, 1208–1216. https://doi.org/10.1016/j.drudis.2012.06.015

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., et al. (2015). The human transcriptome across tissues and individuals. *Science*, *348*(6235), 660–665.

Meyer, P. E., Kontos, K., Lafitte, F., & Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *Eurasip Journal on Bioinformatics and Systems Biology*, *2007*(1). https://doi.org/10.1155/2007/79879

Meyer, P., Marbach, D., Roy, S., & Kellis, M. (2010). Information-Theoretic Inference of Gene Networks Using Backward Elimination. *Conference on Bioinformatics & Computational Biology (BIOCOMP'10)*, 700–705. https://doi.org/10.1093/nar/gks762

Mostafavi, S, Battle, A, Zhu, X, Potash, J., Weissman, M., Shi, J, Beckman, K, Haudenschild, C, McCormick, C, Mei, R, et al. (2014). Type I interferon signaling genes in recurrent major depression: increased expression detected by whole-blood RNA sequencing. *Molecular Psychiatry*, *19*(12), 1267–1274.

Mostafavi, S., Battle, A., Zhu, X., Urban, A. E., Levinson, D., Montgomery, S. B., & Koller, D. (2013). Normalizing RNA-sequencing data by modeling hidden covariates with prior knowledge. *PloS ONE*, *8*(7), e68141.

Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *368*(1620), 20120362. https://doi.org/10.1098/rstb.2012.0362

Ong, C.-T., & Corces, V. G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics*, *12*(4), 283–93.

Pappas, C. T., Mayfield, R. M., Henderson, C., Jamilpour, N., Cover, C., Hernandez, Z., Hutchinson, K. R., Chu, M., Nam, K.-H., Valdez, J. M., Wong, P. K., Granzier, H. L., & Gregorio, C. C. (2015). Knockout of Lmod2 results in shorter thin filaments followed by dilated cardiomyopathy and juvenile lethality. *Proceedings of the National Academy of Sciences*, *112*(44), 13573–13578.

Parsana, P., Ruberman, C., Jaffe, A. E., Schatz, M. C., Battle, A., & Leek, J. T. (2019). Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome Biology*, *20*(1), 94. https://doi.org/10.1186/s13059-019-1700-9

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Penrod, N. M., Cowper-Sal-Lari, R., & Moore, J. H. (2011). Systems genetics for drug target discovery. *Trends in Pharmacological Sciences*, *32*(10), 623–630.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295. https://doi.org/10.1038/nbt.3122

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, *9*(1), 171–181. https://doi.org/10.1038/nprot.2014.006

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., & Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, *464*(7289), 768–772. https://doi.org/10.1038/nature08872

Pierson, E., Koller, D., Battle, A., Mostafavi, S., Consortium, G., et al. (2015). Sharing and specificity of co-expression networks across 35 human tissues. *PLoS Computational Biology*, *11*(5), e1004220.

Pink, R. C., Wicks, K., Caley, D. P., Punch, E. K., Jacobs, L., & Carter, D. R. F. (2011). Pseudogenes: pseudo-functional or key regulators in health and disease? *RNA*, *17*(5), 792–8. https://doi.org/10.1261/rna.2658311

Piro, R. M., Ala, U., Molineris, I., Grassi, E., Bracco, C., Perego, G. P., Provero, P., & Di Cunto, F. (2011). An atlas of tissue-specific conserved coexpression for functional annotation and disease gene prediction. *European Journal of Human Genetics*, *19*(11), 1173–1180.

Prieto, C., Risueño, A., Fontanillo, C., & De Las Rivas, J. (2008). Human gene coexpression landscape: Confident network derived from tissue transcriptomic profiles. *PLoS ONE*, *3*(12).

Qian, J., Esumi, N., Chen, Y., Wang, Q., Chowers, I., & Zack, D. J. (2005). Identification of regulatory targets of tissue-specific transcription factors: application to retina-specific gene regulation. *Nucleic Acids Research*, *33*(11), 3479–3491.

Reilly, C., Raghavan, A., & Bohjanen, P. (2006). Global assessment of cross-hybridization for oligonucleotide arrays. *Journal of biomolecular techniques : JBT*, *17*(2), 163–72.

Robert, C., & Watson, M. (2015). Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome biology*, *16*(1), 177. https://doi.org/10.1186/s13059-015-0734-x

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. https://doi.org/10.1186/gb-2010-11-3-r25

Roider, H. G., Manke, T., O'keeffe, S., Vingron, M., & Haas, S. A. (2009). PASTAA: Identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics*, *25*(4), 435–442.

Saha, A., & Battle, A. (2018). False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors. *F1000Research*, *7*, 1860. https://doi.org/10.12688/f1000research.17145.1

Saha, A., & Battle, A. (2019a). Battle-lab/crossmap: Github repository to compute cross-mappability (release 1.2). https://doi.org/10.5281/zenodo.2602096

Saha, A., & Battle, A. (2019b). Pre-computed cross-mappability resources for human genomes (hg19 and grch38). https://doi.org/10.6084/m9.figshare.c.4297352.v4

Saha, A., Jeon, M., Tan, A. C., & Kang, J. (2015). iCOSSY: An Online Tool for Context-Specific Subnetwork Discovery from Gene Expression Data (C. Romualdi, Ed.). *PLOS ONE*, *10*(7), e0131656. https://doi.org/10.1371/journal.pone.0131656

Saha, A., Kim, Y., Gewirtz, A. D. H., Jo, B., Gao, C., McDowell, I. C., Consortium, T. G., Engelhardt, B. E., & Battle, A. (2017). Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome research*. https://doi.org/10.1101/gr.216721.116

Saha, A., Tan, A. C. A., & Kang, J. (2014). Automatic context-specific subnetwork discovery from large interaction networks. (P. Aloy, Ed.). *PloS ONE*, *9*(1), e84227. https://doi.org/10.1371/journal.pone.0084227

Scotti, M. M., & Swanson, M. S. (2015). RNA mis-splicing in disease. *Nature Reviews Genetics*, *17*(1), 19–32.

Seo, C. H., Kim, J.-R., Kim, M.-S., & Cho, K.-H. (2009). Hub genes with positive feedbacks function as master switches in developmental gene regulatory networks. *Bioinformatics*, *25*(15), 1898–1904. https://doi.org/10.1093/bioinformatics/btp316

Seoane, J. A., Kirkland, J. G., Caswell-Jin, J. L., Crabtree, G. R., & Curtis, C. (2019). Chromatin regulators mediate anthracycline sensitivity in breast cancer. *Nature Medicine*, *25*(11), 1721–1727. https://doi.org/10.1038/s41591-019-0638-5

Shabalin, A. A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, *28*(10), 1353–8.

Squier, C. A., & Kremer, M. J. (2001). Biology of oral mucosa and esophagus. *Journal of the National Cancer Institute. Monographs*, *2001*(29), 7–15.

Stegle, O, Parts, L, Piipari, M, Winn, J, & Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, *7*(3), 500–507.

Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, *302*(5643), 249–255.

Sui, Y, Yang, Z, Xiong, S, Zhang, L, Blanchard, K. L., Peiper, S. C., Dynan, W. S., Tuan, D, & Ko, L. (2007). Gene amplification and associated loss of 5′ regulatory sequences of CoAA in human cancers. *Oncogene*, *26*(6), 822–835.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., & Von Mering, C. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, *47*(D1), D607–D613. https://doi.org/10.1093/nar/gky1131

The GTEx Consortium. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660.

The GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), 204–213. https://doi.org/10.1038/nature24277

The GTEx Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, *369*(6509), 1318–1330. https://doi.org/10.1126/science.aaz1776

Tomsic, J., He, H., Akagi, K., Liyanarachchi, S., Pan, Q., Bertani, B., Nagy, R., Symer, D. E., Blencowe, B. J., & de la Chapelle, A. (2015). A germline mutation in SRRM2, a splicing factor gene, is implicated in papillary thyroid carcinoma predisposition. *Scientific Reports*, *5*, 10566.

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, *25*(9), 1105–1111. https://doi.org/10.1093/bioinformatics/btp120

U.S. Food and Drug Administration. (2020). Table of Pharmacogenomic Biomarkers in Drug Labeling. Retrieved February 9, 2021, from https://www.fda.gov/drugs/science-and-research-drugs/table-pharmacogenomic-biomarkers-drug-labeling

Valverde, P., Healy, E., Jackson, I., Rees, J. L., Thody, A. J., & Victoria, R. (1995). Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. *Nature genetics*, *11*, 328–330.

van de Geijn, B., McVicker, G., Gilad, Y., & Pritchard, J. K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods*, *12*(11), 1061–1063. https://doi.org/10.1038/nmeth.3582

Vareli, K, Frangou-Lazaridis, M, van der Kraan, I, Tsolas, O, & van Driel, R. (2000). Nuclear distribution of prothymosin alpha and parathymosin: evidence that prothymosin alpha is associated with RNA synthesis processing and parathymosin with early DNA replication. *Experimental Cell Research*, *257*(1), 152–61.

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., Pervjakova, N., Alvaes, I., Fave, M.-J., Agbessi, M., Christiansen, M., Jansen, R., Seppälä, I., Tong, L., Teumer, A., … Franke, L. (2018). Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv*, 447367. https://doi.org/10.1101/447367

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., & Burge, C. B. (2008). Alternative

isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–6.

Wang, Z., & Burge, C. B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, *14*(5), 802–13.

Ward, A. J., & Cooper, T. A. (2010). The pathobiology of splicing. *The Journal of Pathology*, *220*(2), 152–163.

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, *38*(Web Server issue), W214–W220.

Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, *171*, 737–738. https://doi.org/10.1038/171737a0

Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M. W., Fairfax, B. P., Schramm, K., Powell, J. E., Zhernakova, A., Zhernakova, D. V., Veldink, J. H., Van den Berg, L. H., Karjalainen, J., Withoff, S., Uitterlinden, A. G., Hofman, A., Rivadeneira, F., … Franke, L. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature genetics*, *45*(10), 1238–43. https://doi.org/10.1038/ng.2756

Witten, J. T., & Ule, J. (2011). Understanding splicing regulation through RNA splicing maps. *Trends in Genetics*, *27*(3), 89–97.

Wu, J. Y., Kar, A., Kuo, D., Yu, B., & Havlioglu, N. (2006). SRp54 (SFRS11), a regulator for tau exon 10 alternative splicing identified by an expression cloning strategy. *Molecular and Cellular Biology*, *26*(18), 6739–47.

Xia, C., Fan, J., Emanuel, G., Hao, J., & Zhuang, X. (2019). Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proceedings of the National Academy of Sciences*, *116*(39), 19490–19499. https://doi.org/10.1073/pnas.1912459116

Xiao, X., Moreno-moral, A., Rotival, M., Bottolo, L., & Petretto, E. (2014). Multitissue Analysis of Co-expression Networks by Higher-Order Generalized Singular Value Decomposition Identifies Functionally Coherent Transcriptional Modules. *PLoS Genetics*, *10*(1).

Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., & Liang, H. (2014). Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature Communications*, *5*, 1–9.

Yin, L., Cai, Z., Zhu, B., & Xu, C. (2018). Identification of Key Pathways and Genes in the Dynamic Progression of HCC Based on WGCNA. *Genes*, *9*(2), 92. https://doi.org/10.3390/genes9020092

Zhai, X., Xue, Q., Liu, Q., Guo, Y., & Chen, Z. (2017). Colon cancer recurrence-associated genes revealed by WGCNA co-expression network analysis. *Molecular Medicine Reports*, *16*(5), 6499–6505. https://doi.org/10.3892/mmr.2017.7412

Zhang, B., & Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*(1). https://doi.org/10.2202/1544-6115.1128

Zhang, W. J., & Wu, J. Y. (1996). Functional properties of p54, a novel SR protein active in constitutive and alternative splicing. *Molecular and Cellular Biology*, *16*(10), 5400–8.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*(1), 1–12. https://doi.org/10.1038/ncomms14049

Zhong, R., Allen, J. D., Xiao, G., & Xie, Y. (2014). Ensemble-based network aggregation improves the accuracy of gene network reconstruction. *PLoS ONE*, *9*(11), 1–10.

Zhou, A., Ou, A. C., Cho, A., Benz, E. J., & Huang, S.-C. (2008). Novel splicing factor RBM25 modulates Bcl-x pre-mRNA 5′ splice site selection. *Molecular and Cellular Biology*, *28*(19), 5924–36.

Zimowska, G., Shi, J., Munguba, G., Jackson, M. R., Alpatov, R., Simmons, M. N., Shi, Y., & Sugrue, S. P. (2003). Pinin/DRS/memA Interacts with SRp75, SRm300 and SRrp130 in Corneal Epithelial Cells. *Investigative Opthalmology & Visual Science*, *44*(11), 4715.

Appendix A

Joint regulation of transcription and alternative splicing (appendix)

GO BP id	Biological process name	Number of tissues
GO:0016070	RNA metabolic process	13
GO:0010467	gene expression	12
GO:0090304	nucleic acid metabolic process	12
GO:0044260	cellular macromolecule metabolic process	11
GO:0006396	RNA processing	9
GO:0006725	cellular aromatic compound metabolic process	8
GO:0008380	RNA splicing	8
GO:0043170	macromolecule metabolic process	8
GO:0044237	cellular metabolic process	8
GO:0046483	heterocycle metabolic process	8
GO:0006139	nucleobase-containing compound metabolic process	7
GO:0006807	nitrogen compound metabolic process	7
GO:0034645	cellular macromolecule biosynthetic process	7
GO:0044238	primary metabolic process	7
GO:0016071	mRNA metabolic process	6
GO:0034641	cellular nitrogen compound metabolic process	6
GO:0071704	organic substance metabolic process	6
GO:1901360	organic cyclic compound metabolic process	6
GO:0006397	mRNA processing	5
GO:0032774	RNA biosynthetic process	5
GO:0051252	regulation of RNA metabolic process	5
GO:2000112	regulation of cellular macromolecule biosynthetic process	5

Table A.1: Top GO biological processes among TE-IR hubs. We tested for enrichment of all GO biological processes (BPs) in top 500 TE-IR hubs. We selected a BP term in our analysis if that had at least 20 genes in our data. This table summarizes the top BP terms based on the number of tissues they appear in top 20 strongest terms (lowest p) in individual tissues.

GO MF id	Molecular function name	Number of tissues
GO:0003723	RNA binding	15
GO:0044822	poly(A) RNA binding	15
GO:0003676	nucleic acid binding	14
GO:1901363	heterocyclic compound binding	14
GO:0003677	DNA binding	13
GO:0097159	organic cyclic compound binding	13
GO:0001071	nucleic acid binding transcription factor activity	9
GO:0003700	sequence-specific DNA binding transcription factor activity	9
GO:0000975	regulatory region DNA binding	7
GO:0000989	transcription factor binding transcription factor activity	7
GO:0001067	regulatory region nucleic acid binding	7
GO:0008168	methyltransferase activity	7
GO:0044212	transcription regulatory region DNA binding	7
GO:0000981	sequence-specific DNA binding RNA polymerase II transcription factor activity	6
GO:0000988	protein binding transcription factor activity	6
GO:0003712	transcription cofactor activity	6
GO:0005488	binding	6
GO:0043565	sequence-specific DNA binding	5
GO:0043566	structure-specific DNA binding	5
GO:0003713	transcription coactivator activity	4
GO:0003714	transcription corepressor activity	4
GO:0016741	transferase activity, transferring one-carbon groups	4

Table A.2: Top GO molecular functions among TE-IR hubs. We tested for enrichment of all GO molecular functions (MFs) in top 500 TE-IR hubs. We selected a MF term in our analysis if that had at least 20 genes in our data. This table summarizes the top MF terms based on the number of tissues they appear in top 20 strongest terms (lowest p) in individual tissues.

Tissue	TE–TE	TE–IR	IR-TE	IR-IR
Adipose – Subcutaneous	62.50%	38.46%	NA	NA
Adipose – Visceral	53.85%	39.13%	100%	100%
Artery – Aorta	56.25%	38.46%	80%	100%
Artery – Tibial	53.85%	42.86%	100%	100%
Breast – Mammary	35.29%	55.17%	100%	100%
Cells – Transformed Fibroblasts	71.88%	55.26%	100%	100%
Esophagus – Mucosa	61.54%	62.50%	100%	100%
Esophagus – Muscularis	72.09%	20.00%	83.33%	85.71%
Heart – Left Ventricle	61.11%	79.49%	100%	100%
Lung	91.43%	30.00%	100%	100%
Muscle – Skeletal	69.70%	80.77%	100%	100%
Nerve – Tibial	69.57%	53.85%	100%	100%
Skin – Not Sun Exposed	70.00%	42.86%	100%	100%
Skin – Sun Exposed	41.67%	43.75%	100%	100%
Thyroid	75.00%	72.73%	100%	100%
Whole Blood	67.44%	79.31%	100%	100%

Table A.3: Differential expression in tissue-specific hubs. Here, we consider a hub (rank ≤ 100) is tissue-specific if it is not present in top 500 hubs of any other tissues. This table shows the percentage of tissue-specific hubs, categorized by hub type, with at least a 1.5 fold expression level change between the tissue of interest and all other tissues. Here, NA means there was no tissue-specific hub for corresponding category.

Tissue	Unique Expression	Differential Expression	Other
Adipose – Subcutaneous	1.49%	66.58%	31.93%
Adipose – Visceral	2.99%	60.98%	36.04%
Artery – Aorta	4.22%	64.08%	31.71%
Artery – Tibial	1.98%	71.88%	26.14%
Breast – Mammary	4.24%	52.92%	42.85%
Cells – Transformed Fibroblasts	17.56%	70.97%	11.47%
Esophagus – Mucosa	6.65%	75.35%	18.01%
Esophagus – Muscularis	3.90%	61.98%	34.13%
Heart – Left Ventricle	14.58%	77.83%	7.59%
Lung	9.89%	61.62%	28.49%
Muscle – Skeletal	8.53%	83.08%	8.40%
Nerve – Tibial	4.13%	74.01%	21.86%
Skin – Not Sun Exposed	3.91%	70.86%	25.23%
Skin – Sun Exposed	2.45%	71.11%	26.44%
Thyroid	5.29%	77.29%	17.41%
Whole Blood	18.61%	77.41%	3.97%

Table A.4: Sources of tissue-specificity of edges. *Unique Expression*: Both nodes connected by an edge were jointly included in TWN reconstruction of the tissue of interest only i.e., at least one of the nodes were excluded in every other tissue due to low expression or other filtering criteria. *Differential Expression*: Both nodes connected by an edge were jointly included in TWN reconstruction of multiple tissues and at least one of the nodes was differentially expressed (at least 1.5 fold change in raw TPM) between the tissue of interest and rest of the tissues. *Other*: Any other source.

Tissue	Variant	trans-eGene	cis-eGene	p-value	FDR
Brain – Frontal Cortex	rs11065155	*COX5B*	*TRIAP1*	0.02	0.09
Brain – Anterior Cingulate Cortex	rs470411	*MAGOH*	*TRIM29*	3.36×10^{-3}	0.02
Brain – Cerebellar Hemisphere	rs66500423	*UQCRQ*	*NUMBL*	2.36×10^{-2}	0.15
Brain – Cerebellar Hemisphere	rs66500423	*UNC50*	*NUMBL*	3.23×10^{-2}	0.15
Brain – Putamen	rs9371531	*CHCHD1*	*RMND1*	6.99×10^{-4}	0.06
Thyroid	rs934937	*BRCA1*	*C15orf52*	0.00299	0.09

Table A.5: Summary of tissue-specific trans-eQTLs from the cis-eQTL enrichment tests in the TSNs. Columns include tissues, the RSID of the associated genetic variant, the trans-eGene, the cis-eGene, the p-value of the trans-eQTL association, and the FDR of this association.

Tissue	Variant	trans-eGene	cis-eGene	p-value	FDR
Pituitary	rs36077494	*PTPRT*	*KIRREL*	4.66×10^{-3}	0.16
Pancreas	rs16913469	*RNF38*	*DDIT4*	4.61×10^{-4}	0.15
Muscle – Skeletal	rs11121453	*SLC7A8*	*NPHP4*	3.54×10^{-4}	0.15
Brain – Substantia Nigra	rs911110	*UQCRQ*	*PCNA*	1.07×10^{-3}	0.10
Brain – Hypothalamus	rs116850387	*LAMTOR2*	*TRIAP1*	1.30×10^{-3}	0.16
Brain – Hypothalamus	rs73221368	*DSCR3*	*TRIAP1*	1.72×10^{-3}	0.16
Brain – Hypothalamus	rs73221368	*LAMTOR2*	*TRIAP1*	1.76×10^{-3}	0.16
Brain – Hypothalamus	rs73216931	*ILK*	*RILPL2*	1.02×10^{-4}	0.16
Brain – Hypothalamus	rs9974252	*TRIAP1*	*DSCR3*	9.24×10^{-4}	0.16
Brain – Hypothalamus	rs28642307	*RILPL2*	*BOLA1*	1.42×10^{-3}	0.16
Brain – Hypothalamus	rs10742976	*LAMTOR2*	*ILK*	4.91×10^{-4}	0.16
Brain – Hypothalamus	rs960177	*RILPL2*	*CALB2*	9.55×10^{-4}	0.16
Brain – Frontal Cortex	rs2347443	*MAGOH*	*HSCB*	1.33×10^{-4}	0.13
Brain – Cortex	rs45567235	*TRIAP1*	*CCDC107*	5.73×10^{-5}	0.09

Table A.6: Tissue-specific trans-eQTLs from the 20 kb tests in the TSNs. Columns include tissues, the RSID of the associated genetic variant, the trans-eGene, the cis-eGene, the p-value of the trans-eQTL association, and the FDR of this association. Only the most significant trans-eVariant per cis-eGene and trans-eGene pair is included in the table.

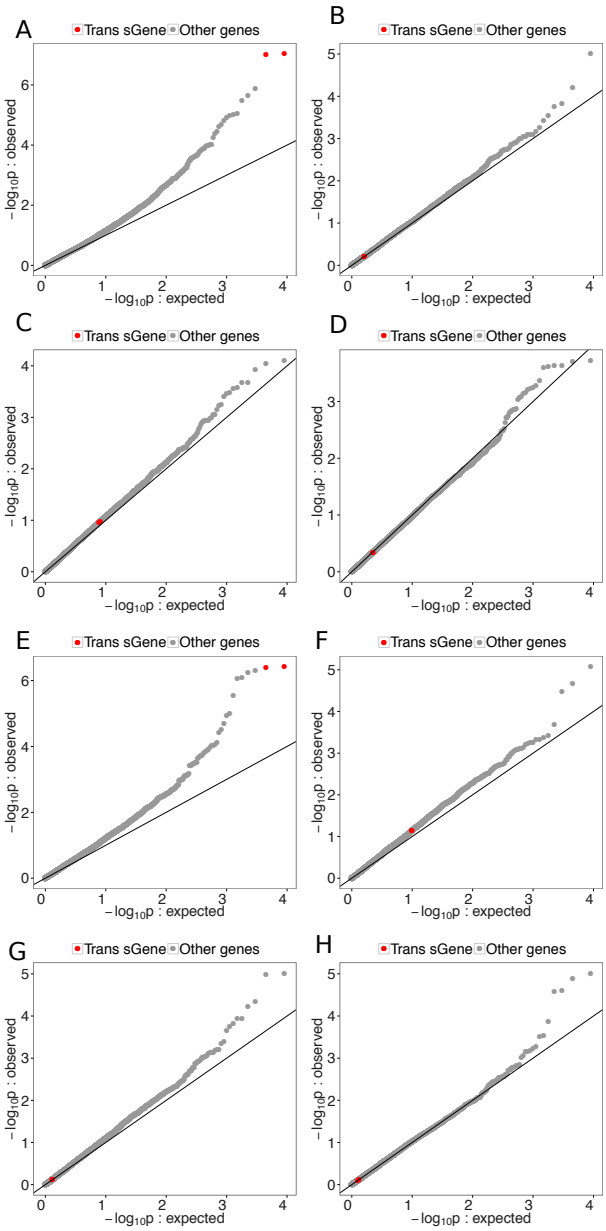

Figure A.1: Association of rs113305055 and rs59153288 with distal isoform ratio across multiple tissues. We measured association for each variant with all isoform ratios genome-wide, and plotted observed p-values against uniformly distributed expected p-values. Top four plots shows enrichment of rs113305055 in *artery – tibial* (A), *whole blood* (B), *skeletal muscle* (C), and *thyroid* (D). Bottom four plots show enrichment of rs59153288 in *breast – mammary*(E), *artery – aorta* (F), *whole blood* (G), and *skin – not sun exposed* (H).

Biography

Ashis is a Ph.D. candidate advised by Dr. Alexis Battle in the Department of Computer Science at Johns Hopkins University (JHU). Before beginning his Ph.D. at JHU, he completed his Master's degree in Computer Engineering from Korea University, South Korea, received his Bachelor's degree in Computer Science and Engineering from Bangladesh University of Engineering and Technology, Bangladesh, and worked for four years in the software industry. The main focus of his research is deciphering biology using computational approaches. After finishing his graduate work, Ashis will join Genentech as a Senior Scientific Researcher.